A Model-Based Method for Minimizing CVaR and Beyond

Si Yi Meng [1 2] **Robert M. Gower** [2]

Abstract

We develop a variant of the stochastic prox-linear method for minimizing the Conditional Value-at-Risk (CVaR) objective. CVaR is a risk measure focused on minimizing worst-case performance, defined as the average of the top quantile of the losses. In machine learning, such a risk measure is useful to train more robust models. Although the stochastic subgradient method (SGM) is a natural choice for minimizing the CVaR objective, we show that our stochastic prox-linear (SPL+) algorithm can better exploit the structure of the objective, while still providing a convenient closed form update. Our SPL+ method also adapts to the scaling of the loss function, which allows for easier tuning. We then specialize a general convergence theorem for SPL+ to our setting, and show that it allows for a wider selection of step sizes compared to SGM. We support this theoretical finding experimentally.

1. Introduction

The most common approach to fit a model parametrized by $\theta \in \mathbb{R}^d$ to data, is to minimize the *expected* loss over the data distribution, that is

$$\min_{\theta \in \mathbb{R}^d} R_{\mathrm{ERM}}(\theta) = \mathbb{E}_{z \sim P}[\ell(\theta; z)]. \tag{1}$$

But in many cases, the expected loss may not be the suitable objective to minimize. When robustness or safety of the model are concerned, the emphasis should rather be on the extreme values of the distribution rather than the average value. For instance, in distributionally robust optimization, the goal is to optimize the model for the worst case distribution around some fixed distribution (Duchi & Namkoong, 2018). In extreme risk-averse settings, such as when safety is the top priority, it is desirable to minimize the maximum loss within a training set (Shalev-Shwartz & Wexler, 2016). These applications can all be formulated as minimizing the expectation of the losses that are *above* some cutoff value,

$$\min_{\theta \in \mathbb{R}^d} R_{\mathrm{CVaR}}(\theta) = \mathbb{E}_{z \sim P}\left[\ell(\theta; z) \mid \ell(\theta; z) \geq \alpha_\beta(\theta)\right], \tag{2}$$

where $\alpha_\beta(\theta)$ is the upper β–quantile of the losses. For example, for $\beta = 0.9$, the problem in (2) is to minimize the expectation of the worst 10% of the losses.

In this work, we propose a variant of the stochastic prox-linear (SPL) method pioneered by Burke & Ferris (1995); Lewis & Wright (2016); Duchi & Ruan (2018) for solving (2). The possibility of applying SPL to CVaR minimization was mentioned in Davis & Drusvyatskiy (2019), but not explored. We introduce a variant of SPL called SPL+, that adapts to the scaling of the loss function, which in turn allows for a default parameter setting. We first derive a closed-form update for SPL+, and show why it is particularly well suited for minimizing CVaR. We give its convergence rates for convex and Lipschitz losses by adapting existing results from Davis & Drusvyatskiy (2019). Through several experiments comparing the stochastic prox-linear method to stochastic subgradient we show that SPL and SPL+ are more robust to the choice of step size. We conclude with a discussion on several future applications for minimizing CVaR in machine learning.

1.1. Background

The CVaR objective was first introduced in finance as an alternative measure of risk, also known as the expected shortfall (Artzner et al., 1999; Embrechts et al., 1999). Many applications in finance can be formulated as CVaR minimization problems, such as portfolio optimization (Krokhmal et al., 2002; Mansini et al., 2007), insurance (Embrechts et al., 2013) and credit risk management (Andersson et al., 2001). The seminal work of Rockafellar & Uryasev (2000) proposed a variational formulation of the CVaR objective that is amenable to standard optimization methods. This formulation has since inspired considerable research in applications spanning machine learning and adjacent fields, such as ν-SVM (Takeda & Sugiyama, 2008; Gotoh & Takeda, 2016), robust decision making and MDPs (Chow et al., 2015; Chow & Ghavamzadeh, 2014; Chow et al., 2017; Cardoso & Xu, 2019; Sani et al., 2012), influence maximization and submodular optimization (Maehara, 2015; Ohsaka & Yoshida,

[1]Department of Computer Science, Cornell University, Ithaca, NY, USA [2]Center for Computational Mathmatics, Flatiron Institute, New York, NY, USA. Correspondence to: Si Yi Meng <sm2833@cornell.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

2017; Wilder, 2018), fairness (Williamson & Menon, 2019), and federated learning (Laguel et al., 2021b).

Though it finds many applications, the CVaR objective is typically difficult to minimize. It is nonsmooth even when the individual losses $\ell(\cdot; z)$ are continuously differentiable. Indeed, if P does not admit a density — which is the case for all empirical distributions over training data — the variational objective is not everywhere differentiable. To address this, Laguel et al. (2021a) developed subdifferential calculus for a number of equivalent CVaR formulations and proposed minimizing a smoothed version of the dual objective. On the other hand, several works (Soma & Yoshida, 2020; Holland & Haress, 2021) apply the stochastic subgradient method directly to the variational formulation proposed by Rockafellar & Uryasev (2000), which is well-defined regardless of the distribution P. However, as we elaborate in Section 3, this approach is oblivious to the special structure of the variational form of the CVaR objective.

2. Problem setup

Let $\ell(\theta; z)$ be the loss associated with the model parameters $\theta \in \mathbb{R}^d$ and a measurable random variable $z(\omega)$ on some background probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

When z follows a distribution P with density $p(z)$, the cumulative distribution function on the loss for a fixed θ is given by $\mathbb{P}[\ell(\theta; z) \leq \alpha] = \int_{\ell(\theta;z)\leq\alpha} p(z)\, dz$, which we assume is everywhere continuous with respect to α. Let β be a confidence level, for instance $\beta = 0.9$. The Value-at-Risk (VaR) of the model is the lowest α such that with probability β, the loss will not exceed α. Formally,

$$\mathrm{VaR}_\beta(\theta) := \min\{\alpha \in \mathbb{R} \,:\, \mathbb{P}[\ell(\theta; z) \leq \alpha] \geq \beta\}. \quad (3)$$

The Conditional Value-at-Risk (CVaR) is the expectation of the upper tail starting at VaR_β, illustrated in Figure 1:

$$\mathrm{CVaR}_\beta(\theta) := \mathbb{E}_{z\sim P}[\ell(\theta; z) \mid \ell(\theta; z) \geq \mathrm{VaR}_\beta(\theta)]. \quad (4)$$

Clearly, the CVaR upper bounds the VaR for the same β. Our goal is to minimize CVaR_β over $\theta \in \mathbb{R}^d$, but directly minimizing (4) is not straightforward. Fortunately, Rockafellar & Uryasev (2000) introduced a variational formulation where the solution to

$$\theta^*, \alpha^* \in \operatorname*{arg\,min}_{\theta\in\mathbb{R}^d, \alpha\in\mathbb{R}} F_\beta(\theta, \alpha) \quad \text{where,} \quad (5)$$

$$F_\beta(\theta, \alpha) := \alpha + \frac{1}{1-\beta}\mathbb{E}_{z\sim P}\left[\max\{\ell(\theta; z) - \alpha,\, 0\}\right]$$

is such that θ^* is the solution to (4), and we obtain $\alpha^* = \mathrm{VaR}_\beta(\theta)$ as a byproduct.

3. The Stochastic Subgradient Method

A natural choice for minimizing (5) is the stochastic subgradient method (SGM). Letting ∂f denote the convex subdifferential of f, at each step t we sample $z \sim P$ uniformly and compute a subgradient g_t from the subdifferential

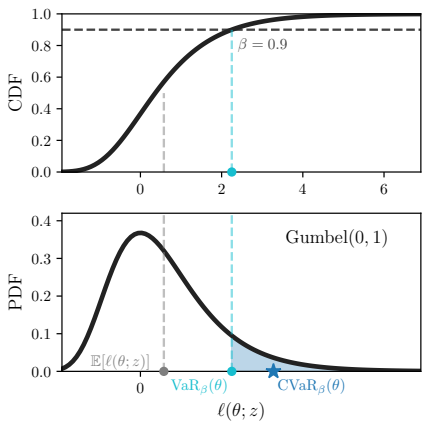

Figure 1: Expectation, VaR, and CVaR.

$$\partial F_\beta(\theta_t, \alpha_t; z) = \begin{pmatrix}\mathbf{0}\\ 1\end{pmatrix} + \frac{u_t}{1-\beta}\begin{pmatrix}\partial\ell(\theta_t; z)\\ -1\end{pmatrix} \quad (6)$$

where $u_t = \partial \max\{u, 0\}|_{u\,=\,\ell(\theta_t;\,z)\,-\,\alpha_t}$. Given some step size sequence $\{\lambda_t\} > 0$, and denoting $x = (\theta, \alpha)^\intercal$, SGM then takes the step

$$x_{t+1} = x_t - \lambda_t g_t, \quad \text{where } g_t \in \partial F_\beta(\theta_t, \alpha_t; z). \quad (7)$$

Substituting in the subgradient g_t given in (6) into (7) gives

$$\theta_{t+1} = \theta_t - \frac{\lambda_t}{1-\beta} u_t \partial\ell(\theta_t; z), \quad (8)$$

$$\alpha_{t+1} = \alpha_t - \lambda_t + \frac{\lambda_t}{1-\beta} u_t, \quad (9)$$

For reference, the complete SGM algorithm is given in Algorithm 1. SGM is very sensitive to the step size choice and may diverge if not carefully tuned. This issue can be explained from a modeling perspective (Davis & Drusvyatskiy, 2019). Indeed, SGM can be written as a model-based method where at each iteration t, it uses the following linearization of the sampled $F_\beta(x; z)$ at the current point x_t:

$$m_t^{\mathrm{SGM}}(x; z) := F_\beta(x_t; z) + \langle g_t,\, x - x_t\rangle. \quad (10)$$

This provides an approximate, stochastic model of the objective $F_\beta(x)$. The SGM update is then a proximal step on this model, that is

$$x_{t+1} = \operatorname*{arg\,min}_{x\in\mathbb{R}^{d+1}} m_t(x; z) + \frac{1}{2\lambda_t}\|x - x_t\|^2 \quad (11)$$

using $m_t = m_t^{\mathrm{SGM}}$. The issue with $m_t = m_t^{\mathrm{SGM}}(x; z)$ is that it uses a linearization to approximate the $\max\{\cdot, 0\}$ function. This linearization can take negative values, which

Algorithm 1 SGM: Stochastic subgradient method for CVaR minimization

initialize: $\theta_0 \in \mathbb{R}^d$, $\alpha_0 \in \mathbb{R}$, **hyperparameter:** $\lambda > 0$
for $t = 0, 1, 2, \ldots, T$ **do**
  Sample data point $z \sim P$, compute $\ell(\theta_t; z)$ and $v_t \in \partial \ell(\theta_t; z)$
  $\lambda_t \leftarrow \lambda / \sqrt{t+1}$
  if $\alpha_t \geq \ell(\theta_t; z)$ **then** ▷ α_t too big
    $\theta_{t+1} \leftarrow \theta_t$
    $\alpha_{t+1} \leftarrow \alpha_t - \lambda_t$
  else ▷ α_t too small
    $\theta_{t+1} \leftarrow \theta_t - \frac{\lambda_t}{1-\beta} v_t$
    $\alpha_{t+1} \leftarrow \alpha_t + \frac{\lambda_t}{1-\beta} \beta$
  end if
end for
return $\bar{x}_T = \frac{1}{T+1} \sum_{t=1}^{T+1} (\theta_t, \alpha_t)^\intercal$

is a poor approximation of the non-negative $\max\{\cdot, 0\}$ operation. The main insight of the SPL method is to leverage the structure of $F_\beta(x)$ as a truncated function. This structure allows for a more accurate model that still has an easily computable proximal operator.

4. The SPL method for CVaR minimization

4.1. A tighter model

Here we introduce an alternative model for our objective that only linearizes *inside* the $\max\{\cdot, 0\}$, which is a strictly more accurate model when the objective is convex (Asi & Duchi, 2019a). In particular, for some $v_t \in \partial \ell(\theta_t; z)$ and $\ell_t := \ell(\theta_t; z)$, we use

$$m_t^{\mathrm{SPL}}(x; z) = \alpha + \frac{\max\left\{\ell_t + \langle v_t, \theta - \theta_t \rangle - \alpha, 0\right\}}{1 - \beta} \quad (12)$$

The algorithm resulting from (11) using $m_t = m_t^{\mathrm{SPL}}$ is known as the stochastic prox-linear (SPL) method (Duchi & Ruan, 2018). Figure 2 illustrates that (12) better approximates the level sets of the loss function as compared to (10).

4.2. Separate regularization parameters

Now that we have determined a tighter model (12), it remains now to select a default step size sequence λ_t for the proximal step (11). But, as we will argue next, having the same default step size sequence for both α and θ could lead to inconsistencies due to the dependency on the *scale* of the loss function.

To explain this dependency, let $\mathrm{units}(\ell)$ denote the *units* of our loss function $\ell(\theta_t; z)$. For instance, our loss could be a cost measured in dollars. Since α approximates a quantile of the losses, it must also have the same units as the loss.

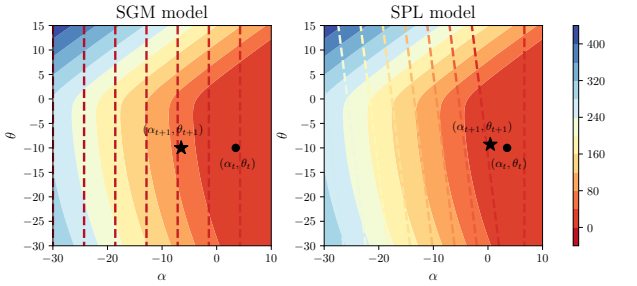

Figure 2: Comparison of SGM and SPL models on the CVaR objective with a single $\ell(\theta) = \log(1 + \exp(\theta)) + \frac{0.01}{2}\theta^2$. Filled contours are the level sets of the objective, while the dashed contour lines are the level sets of the respective model m_t constructed at (θ_t, α_t). With the same step size, the SGM model results in an update that increases the objective, whereas the SPL model does not. Note that because the subgradient of the objective is 0 in θ, the SGM model is constant in θ.

Consequently, our model in (12) also has the same units as the loss function. A clash of units appears when we consider the regularization term in (11), that is the term

$$\frac{1}{2\lambda_t} \|x - x_t\|^2 = \frac{1}{2\lambda_t} \left(\|\theta - \theta_t\|^2 + (\alpha - \alpha_t)^2 \right).$$

This regularization term must also have the same units as the loss so that the entire objective in (11) has consistent units. But since $\mathrm{units}(\alpha) = \mathrm{units}(\ell)$, the term $\frac{1}{2\lambda_t}(\alpha - \alpha_t)^2$ can only have the same units as the loss if $\mathrm{units}(\lambda_t) = \mathrm{units}(\ell)$. In direct contradiction, the term $\frac{1}{2\lambda_t} \|\theta - \theta_t\|^2$ can only have the same units as the loss if $\mathrm{units}(\lambda_t) = 1/\mathrm{units}(\ell)$, since θ parametrizes the objective and thus does not carry the units of the loss. There is no choice of λ_t which would result in the objective of (11) having consistent units; consequently, there is no default, scale-invariant λ_t that would work across different loss functions.

One simple way to fix this clash of units, is to disentangle λ_t into two regularization parameters $\lambda_{\theta,t}, \lambda_{\alpha,t} > 0$ and update the iterates according to

$$\theta_{t+1}, \alpha_{t+1} = \operatorname*{arg\,min}_{\theta \in \mathbb{R}^d, \alpha \in \mathbb{R}} m_t^{\mathrm{SPL}}(x; z) + \frac{1}{2\lambda_{\theta,t}} \|\theta - \theta_t\|^2 + \frac{1}{2\lambda_{\alpha,t}} (\alpha_t - \alpha)^2. \quad (13)$$

Now we can make the units match across (13) by choosing

$$\mathrm{units}(\lambda_{\alpha,t}) = \mathrm{units}(\ell) \text{ and } \mathrm{units}(\lambda_{\theta,t}) = \frac{1}{\mathrm{units}(\ell)}. \quad (14)$$

As suggested by our theory in (26), if we had access to the average Lipschitz constant L of the the individual losses ℓ, then we should choose

$$\lambda_{\alpha,t} = \frac{\lambda |\alpha_t - \alpha^*|}{\sqrt{t}} \quad \text{and} \quad \lambda_{\theta,t} = \frac{\lambda \|\theta_t - \theta^*\|}{L\sqrt{t}}, \quad (15)$$

where $\lambda > 0$ is a numerical constant. Although this gives us consistency in the units, estimating L can be difficult in practice. Thus, instead we approximate the scaling by using the initial loss $\ell_0 := \mathbb{E}_z[\ell(\theta_0; z)]$ and choose

$$\lambda_{\alpha,t} = \frac{\lambda \ell_0}{\sqrt{t}} \quad \text{and} \quad \lambda_{\theta,t} = \frac{\lambda}{\ell_0 \sqrt{t}}, \tag{16}$$

while setting λ using a grid search. We will use (16) as our default setting for $\lambda_{\theta,t}$ and $\lambda_{\alpha,t}$. Importantly, although we have separate regularization terms, there is still only one hyperparameter λ to be set.

4.3. Closed form update

Lemma 1 (Closed form updates of SPL+)**.** *The closed form solution to* (13) *is given by the updates*

$$\theta_{t+1} = \theta_t - \lambda_{\theta,t} \min\left\{\frac{1}{1-\beta}, \gamma_t\right\} \nabla \ell(\theta_t; z), \tag{17}$$

$$\alpha_{t+1} = \alpha_t - \lambda_{\alpha,t} + \lambda_{\alpha,t} \min\left\{\frac{1}{1-\beta}, \gamma_t\right\}, \tag{18}$$

$$\textit{where} \quad \gamma_t = \frac{\max\{\ell(\theta_t; z) - \alpha_t + \lambda_{\alpha,t}, 0\}}{\lambda_{\theta,t} \|\nabla \ell(\theta_t; z)\|^2 + \lambda_{\alpha,t}}. \tag{19}$$

We first give a sketch of how the updates are derived.

Proof. For one step update, we can drop the subscript t without loss of generality. The key step is to rewrite (13) in the form of a proximal step on a truncated model, namely,

$$x_{t+1} = \operatorname*{arg\,min}_{x \in \mathbb{R}^{d+1}} \max\{c + \langle a, x - x_t \rangle, 0\} + \frac{1}{2\lambda} \|x - x_t\|^2$$

where $x = (\theta, \hat{\alpha})^\top$ is the concatenation of θ and a scaled version of α. The solution to this has a nice form given in Lemma 2 in the appendix,

$$x^{t+1} = x_t - \underbrace{\min\left\{\lambda, \frac{\max\{c, 0\}}{\|a\|^2}\right\}}_{=:\eta} a.$$

One can show that by redefining variables as

$$\hat{\alpha} = \sqrt{\frac{\lambda_\theta}{\lambda_\alpha}} \alpha \qquad \text{and} \qquad \hat{\alpha}_t = \sqrt{\frac{\lambda_\theta}{\lambda_\alpha}} \alpha_t - \sqrt{\lambda_\theta \lambda_\alpha},$$

we can absorb the leading α in the model (12) into its regularization term, giving us

$$\alpha + \frac{1}{2\lambda_\alpha}(\alpha - \alpha_t)^2 = \frac{1}{2\lambda_\theta}(\hat{\alpha} - \hat{\alpha}_t)^2 + \text{Const.}\,.$$

After some simple manipulation on the linearization term of (13), we get that

$$c = \frac{1}{1-\beta}\left(\ell(\theta_t; z) - \sqrt{\frac{\lambda_\alpha}{\lambda_\theta}} \hat{\alpha}_t\right), \; a = \frac{1}{1-\beta}\begin{pmatrix} \nabla \ell(\theta_t; z) \\ -\sqrt{\frac{\lambda_\alpha}{\lambda_\theta}}. \end{pmatrix}.$$

Plugging a, c into the update of the truncated model above,

$$\eta = \min\left\{\lambda_\theta, \frac{\max\left\{\ell(\theta_t; z) - \sqrt{\frac{\lambda_\alpha}{\lambda_\theta}} \hat{\alpha}_t, 0\right\}}{\frac{1}{(1-\beta)}(\|\nabla \ell(\theta_t; z)\|^2 + \frac{\lambda_\alpha}{\lambda_\theta})}\right\}.$$

Substituting out $\hat{\alpha}_t$ for α_t and multiplying by a gives us the desired θ_{t+1} and α_{t+1}. □

The detailed proof can be found in Appendix A, with a breakdown of the updates in Algorithm 2. Alternative to our technique, one can also derive these updates by enumerating the KKT conditions after formulating (13) as a constrained minimization problem with an additional slack variable.

Examining the update in Lemma 1, we can see that the cost of computing each iteration of SPL+ is of the same order as computing an iteration of SGM. Finally, if we set the regularization parameters according to the guide in (14), we can see by examining the units of SPL+ that γ_t in (19) is *unitless*. As a result, the units are consistent across the updates of both θ in (17) and α in (18). Next, we discuss two applications of SPL+ which correspond to two extreme settings for the CVaR objective.

Algorithm 2 SPL+: Stochastic prox-linear method for CVaR minimization with separate regularization

initialize: $\theta_0 \in \mathbb{R}^d$, $\alpha_0 \in \mathbb{R}$, **hyperparameter:** $\lambda > 0$
for $t = 0, 1, 2, \ldots, T$ **do**
  Sample data point $z \sim P$
  Compute $\ell(\theta_t; z)$ and $v_t \in \partial \ell(\theta_t; z)$
  $\lambda_{\theta,t} \leftarrow \lambda / (\ell_0 \sqrt{t+1})$
  $\lambda_{\alpha,t} \leftarrow \lambda \ell_0 / \sqrt{t+1}$
  if $\alpha_t > \ell(\theta_t; z) + \lambda_{\alpha,t}$ **then** ▷ α_t too big
    $\theta_{t+1} \leftarrow \theta_t$
    $\alpha_{t+1} \leftarrow \alpha_t - \lambda_{\alpha,t}$
  else if $\alpha_t < \ell(\theta_t; z) - \frac{\lambda_{\theta,t}}{1-\beta}\|v_t\|^2 - \frac{\lambda_{\alpha,t}\beta}{1-\beta}$ **then** ▷ α_t too small
    $\theta_{t+1} \leftarrow \theta_t - \frac{\lambda_{\theta,t}}{1-\beta} v_t$
    $\alpha_{t+1} \leftarrow \alpha_t + \frac{\lambda_{\alpha,t}}{1-\beta}\beta$
  else ▷ α_t in middle range
    $\nu \leftarrow \frac{\ell(\theta_t; z) + \lambda_{\alpha,t} - \alpha_t}{\lambda_{\theta,t}\|v_t\|^2 + \lambda_{\alpha,t}}$
    $\theta_{t+1} \leftarrow \theta_t - \lambda_{\theta,t} \nu \nabla \ell(\theta_t; z)$
    $\alpha_{t+1} \leftarrow \alpha_t - \lambda_{\alpha,t} + \lambda_{\alpha,t} \nu$
  end if
end for
return $\bar{x}_T = \frac{1}{T+1} \sum_{t=1}^{T+1} (\theta_t, \alpha_t)^\top$

4.4. Solving the max loss problem

The SPL+ method can been seen as an extension of recent class of adaptive methods (Gower et al., 2022) for minimizing the max loss, as we detail next. If P is the empirical

distribution over n training examples, setting $\beta = {}^{n-1}/_{n}$ turns the CVaR minimization problem into the max loss minimization problem

$$\min_{\theta\in\mathbb{R}^d} f(\theta) = \max_{i=1,\dots,n} \ell(\theta; z_i). \tag{20}$$

Indeed, if $\beta = {}^{n-1}/_{n}$ then the Value-at-Risk (3) would have to be the max loss, that is, $\alpha = \max_{i=1,\dots,n} \ell(\theta; z_i)$. Plugging this into (5) we have that the second term in $F_\beta(\theta, \alpha)$ is zero, leaving only $F_\beta(\theta, \alpha) = \alpha = \max_{i=1,\dots,n} \ell(\theta; z_i)$.

The max loss problem is an interesting problem in its own right (Shalev-Shwartz & Wexler, 2016). Recently Gower et al. (2022) proposed the *Polyak with slack* methods for solving (20). Our SPL+ improves upon the Polyak with slack methods in two ways: first, SPL+ can be applied to minimizing CVaR for any β, and not just the max loss problem; second, SPL+ can enjoy a default parameter setting due to the two regularization parameters and the consideration around units in (14).

Finally we show that in this setting SPL+ can also been seen as a stochastic algorithm that minimizes the Lagrangian of a slack formulation of Equation (20), where the Lagrange multiplier is equal to ${}^{1}/_{1-\beta}$. We establish this equivalence in Appendix D.

4.5. Solving ERM

When P is the empirical distribution over n training examples, and if $\beta = \frac{1}{n}$, then minimizing the CVaR objective in (5) is equivalent to minimizing the expected risk. This is because $\alpha = \min_{i=1,\dots,n} \ell(\theta, z_i)$ due to (3), and consequently from (4) we have that

$$\begin{aligned}\mathrm{CVaR}_\beta(\theta) &= \mathbb{E}_{z\sim P}[\ell(\theta; z) \mid \ell(\theta; z) \geq \min_{i=1,\dots,n} \ell(\theta, z_i)] \\ &= \mathbb{E}_{z\sim P}[\ell(\theta; z)].\end{aligned}$$

Thus minimizing (5) is equivalent to minimizing the expected risk. As a consequence, SPL+ can also be used as an adaptive method for minimizing the expected risk.

5. Convergence theory

We instantiate the convergence analyses from Davis & Drusvyatskiy (2019) in the case of CVaR minimization, and compare the rates for SGM and SPL+ for losses satisfying the following Assumption.

Assumption 5.1 (Convex, subdifferentiable, and Lipschitz)**.** There exist square integrable random variables $M : \Omega \to \mathbb{R}$ such that for a.e. $z \in \Omega$ and all $\theta \in \mathbb{R}^d$, the sample losses $\ell(\theta; z)$ are convex, subdifferentiable[1], and $M(z)$-Lipschitz.

[1]Historically, the prox-linear method was proposed for composite optimization problems where the inner function is C^1 (Burke & Ferris, 1995). Here we slightly abuse the terminology and allow for general subdifferentiable losses $\ell(\cdot; z)$.

Theorem 5.2 (Convergence rates of SGM and SPL+)**.** *Suppose Assumption 5.1 holds. Let $x^* = (\theta^*, \alpha^*)^\intercal$ be a minimizer of $F_\beta(\theta, \alpha)$, and $x_0 \in \mathbb{R}^d$ an arbitrary initialization. Let $(x_t)_{t=0}^T$ be the iterates given by SGM or SPL+, and $\bar{x}_T = \frac{1}{T+1}\sum_{t=1}^{T+1} x_t$ be the averaged iterate.*

SGM. *If $\lambda_t = \frac{\lambda}{\sqrt{T+1}}$ then the iterates (x_t) given by SGM in (7) satisfy*

$$\begin{aligned}&\mathbb{E}\left[F_\beta(\bar{x}_T) - F_\beta(x^*)\right] \\ &\quad\leq \frac{1}{2}\frac{\|\theta_0 - \theta^*\|^2}{\lambda\sqrt{T+1}} + \frac{1}{2}\frac{(\alpha_0 - \alpha^*)^2}{\lambda\sqrt{T+1}} + \frac{\lambda \mathsf{L}_{\mathrm{SGM}}^2}{\sqrt{T+1}},\end{aligned} \tag{21}$$

where

$$\mathsf{L}_{\mathrm{SGM}}^2 = \mathbb{E}_z\left[\frac{M(z)^2 + 1}{(1-\beta)^2} + 1\right] \tag{22}$$

SPL+. *If $\lambda_{\alpha,t} = \frac{\lambda_\alpha}{\sqrt{T+1}}$ and $\lambda_{\theta,t} = \frac{\lambda_\theta}{\sqrt{T+1}}$, then the iterates (x_t) given by SPL+ given in Lemma 1 satisfy*

$$\begin{aligned}&\mathbb{E}\left[F_\beta(\bar{x}_T) - F_\beta(x^*)\right] \\ &\quad\leq \frac{1}{2}\frac{\|\theta_0 - \theta^*\|^2}{\lambda_\theta\sqrt{T+1}} + \frac{1}{2}\frac{(\alpha_0 - \alpha^*)^2}{\lambda_\alpha\sqrt{T+1}} + \frac{\lambda_\alpha \mathsf{L}_{\mathrm{SPL}}^2}{\sqrt{T+1}},\end{aligned} \tag{23}$$

where

$$\mathsf{L}_{\mathrm{SPL+}}^2 = \mathbb{E}_z\left[\frac{\frac{\lambda_\theta}{\lambda_\alpha}M(z)^2 + 1}{(1-\beta)^2}\right]. \tag{24}$$

This result follows by adapting Theorem 4.4 in Davis & Drusvyatskiy (2019), and we verify the assumptions necessary in Appendix B. In particular, the best bound achieved by SGM via minimizing in λ the RHS of (21) is with

$$\lambda = \frac{\|x_0 - x^*\|}{\mathsf{L}_{\mathrm{SGM}}\sqrt{2}} \tag{25}$$

yielding the rate

$$\mathbb{E}\left[F_\beta(\bar{x}_T) - F_\beta(x^*)\right] \leq \frac{\sqrt{2}\,\|x_0 - x^*\|\,\mathsf{L}_{\mathrm{SGM}}}{\sqrt{T+1}}.$$

Similarly, for SPL+, the best bound is achieved at

$$\lambda_\alpha = \frac{|\alpha_0 - \alpha^*|\,(1-\beta)}{\sqrt{2}}, \quad \lambda_\theta = \frac{\|\theta_0 - \theta^*\|\,(1-\beta)}{\sqrt{2}\,\mathbb{E}_z[M(z)]}, \tag{26}$$

giving us the rate

$$\begin{aligned}\mathbb{E}\left[F_\beta(\bar{x}_T) - F_\beta(x^*)\right] &\leq \frac{\|\theta_0 - \theta^*\|\,\mathbb{E}_z[M(z)] + |\alpha_0 - \alpha^*|}{\sqrt{2}(1-\beta)\sqrt{T+1}} \\ &+ \frac{\|\theta_0 - \theta^*\|\,\mathbb{E}_z[M(z)^2]/\mathbb{E}_z[M(z)] + |\alpha_0 - \alpha^*|}{\sqrt{2}(1-\beta)\sqrt{T+1}}.\end{aligned}$$

We can now use Theorem 5.2 to directly compare the convergence rate of SGM in (21) and SPL+ in (23). First, both methods converge at the $O\left(1/\sqrt{T+1}\right)$ rate. The main difference is in the constants. To ease the comparison, let $\lambda_\alpha = \lambda_\theta = \lambda$. In this case, we can see that the Lipschitz constant of SGM in (22) is always greater than the Lipschitz constant of SPL+ in (24), thus SPL+ has a better constant in its rate of convergence. This is another way to confirm that SPL+ uses a better model of the objective function as compared to SGM. Yet another advantage of SPL+ is the flexibility of having two regularization parameters λ_θ and λ_α, which allows for a method that is independent of the units of the loss.

6. Experiments

We design several experiments to compare, and test the sensitivity of SGM, SPL with only one regularization, that is the updates in Lemma 1 where $\lambda_{\theta,t} = \lambda_{\alpha,t} = \lambda_t$, and our proposed SPL+ updates.

6.1. Synthetic data

First we study the sensitivity of the methods to choices of λ when minimizing the CVaR objective (5). We use three different synthetic distributions, similar to the setup of Holland & Haress (2021), where we experiment various combinations of loss functions $\ell(\cdot; z)$ and data distributions controlled by noise ζ (Table 2). For all problems we set the dimension to be $d = 10$. For regression problems, $\theta_{\text{gen}} \sim \mathcal{U}([0,1]^d)$,

and for classification (logistic regression) we use $\theta_{\text{gen}} \sim \mathcal{U}([0,10]^d)$ to increase linear separability. The loss functions and target generation schemes are listed in Table 2. Each target of the corresponding problem contains an error ϵ from one of the distributions in Table 1, which controls the difficulty level of the problem.

Distribution of ζ	**Parameters**
$\text{Normal}(\mu, \sigma^2)$	$\mu = 0, \sigma = 2$
$\text{Gumbel}(\mu, \beta)$	$\mu = 0, \beta = 4$
$\text{LogNormal}(\mu, \sigma^2)$	$\mu = 2, \sigma = 1$

Table 1: Error distributions in 1D.

Since the expectation in the CVaR objective (5) is difficult to compute in closed form, we evaluate the suboptimality gaps using an empirical average over $N = 10^6$ data points sampled i.i.d. from the corresponding distribution under a single fixed seed. This is done for each error distribution and loss function combination, each giving us the discretization

$$\tilde{F}_\beta(\theta, \alpha) = \alpha + \frac{1}{1-\beta}\frac{1}{N}\sum_{i=1}^{N} \max\left\{\ell(\theta; z_i) - \alpha,\, 0\right\}. \quad (27)$$

We set $\beta = 0.95$ for all experiments, and thus have omitted β from all plot descriptions. We run full-batch L-BFGS to obtain the optimal values for comparison, recorded as θ^*, α^*, and $F^* := \tilde{F}_\beta(\theta^*, \alpha^*)$. For initialization, we set $\alpha_0 \sim \mathcal{U}(0,1)$ and $\theta_0 \sim \mathcal{N}(0, I_d)$ at initialization for all algorithms we compare. They are run for $T = 100,000$ iterations using 5 different seeds that control the randomness of initialization and sampling during the course of optimization. In the sensitivity plots (Figures 3 and 8), solid lines show the median values, while the shaded regions indicate the range over the random seeds. All objective evaluations are on $\tilde{F}_\beta(\bar{\theta}_t, \bar{\alpha}_t)$ using the averaged iterates.

We employ a decreasing step size $\lambda_t = \lambda/\sqrt{t+1}$ for SGM and SPL, while $\lambda_{t,\alpha} = \lambda\ell_0/\sqrt{t+1}$ and $\lambda_{t,\theta} = \lambda/\ell_0\sqrt{t+1}$ for SPL+. We study the sensitivity of the methods to λ, varied over a logarithmically-spaced grid $10^{-6}, 10^{-5}, \ldots, 10^4$, densified around $\lambda = 1$ using the extra grid $10^{-1.5}, 10^{-0.5}, \ldots, 10^{1.5}$.

Figure 3 shows the final suboptimality achieved by SGM, SPL, and SPL+ for different values of λ. For smooth losses (squared and logistic) we see that SPL and SPL+ are significantly more robust and admit a much larger range of λ for which they achieve a low suboptimality. Interestingly, for the absolute loss, the difference is barely noticeable. We also observe that SPL+ often admits a wider basin of good settings for λ as compared to SGM and even SPL. Moreover, $\lambda = 1$ is often in the set of good parameter choices for SPL+. This suggest that our scaling of $\lambda\ell_0$ and λ/θ_0, as motivated by balancing units, lead to a more stable and easy to tune method by choosing λ around 1. In Figure 8, we perform the sensitivity analysis under a fixed accuracy target $\tilde{F}(\theta, \alpha) - \tilde{F}^* \leq \epsilon$, and draw similar stability conclusions.

6.2. Real data

Finally, we present the same experiment on four real datasets: `YearPredictionMSD`, `E2006-tfidf`, (binary) `mushrooms` and (binary) `Covertype`, all from the LIBSVM repository (Chang & Lin, 2011). Similar to the synthetic experiments, we set $\beta = 0.95$ and compute θ^* and α^* using L-BFGS. The objective is now by default the empirical CVaR in (27) since P is the empirical distribution,

$$F_\beta(\theta, \alpha) = \alpha + \frac{1}{1-\beta}\frac{1}{n}\sum_{i=1}^{n} \max\left\{\ell(\theta; z_i) - \alpha,\, 0\right\}$$

where n is the number of examples in the training split. The loss function $\ell(\cdot; z_i)$ is the squared loss for `YearPredictionMSD` and `E2006-tfidf`, and logistic loss for `mushrooms` and `Covertype`. For the comparison between SGM, SPL, and SPL+, we run the methods for $200N$ iterations (except on `E2006-tfidf` where we only run for $10N$ iterations due to its size). All convergence

Table 2: Loss functions and data generation used for synthetic problems. The error distributions for ζ are described in Table 1. We use $\sigma(\cdot)$ to denote the sigmoid function, and all x's are sampled uniformly from the unit sphere.

Task	**Loss** $\ell(\theta; x, y)$	**Target**
Regression	$\frac{1}{2}(x^\intercal\theta - y)^2$	$y = x^\intercal\theta_{\text{gen}} + \zeta$
Regression	$\lvert x^\intercal\theta - y\rvert$	$y = x^\intercal\theta_{\text{gen}} + \zeta$
Classification	$\log\left(1 + \exp\left(-yx^\intercal\theta\right)\right)$	$y = 1$ w.p. $\sigma(x^\intercal\theta_{\text{gen}} + \zeta)$ and -1 otherwise.

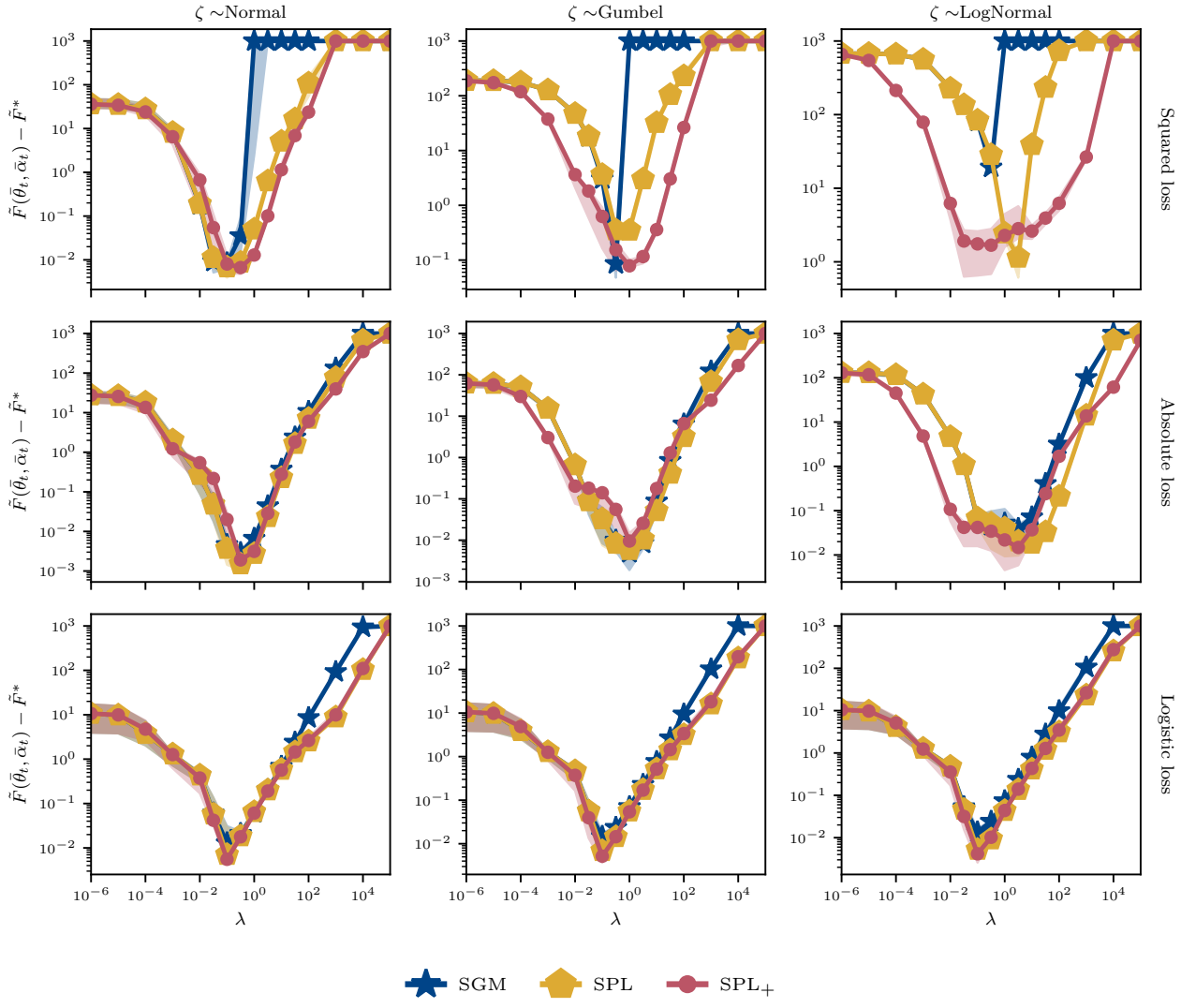

Figure 3: Sensitivity of final suboptimality to step size choices under a fixed $T = 10^5$ budget. The first two rows are regression tasks under the ℓ_1 and ℓ_2 losses, while the third row correspond to a binary classification task under the logistic loss. The columns correspond to different noise distributions in the data generation that controls the difficulty of the problem.

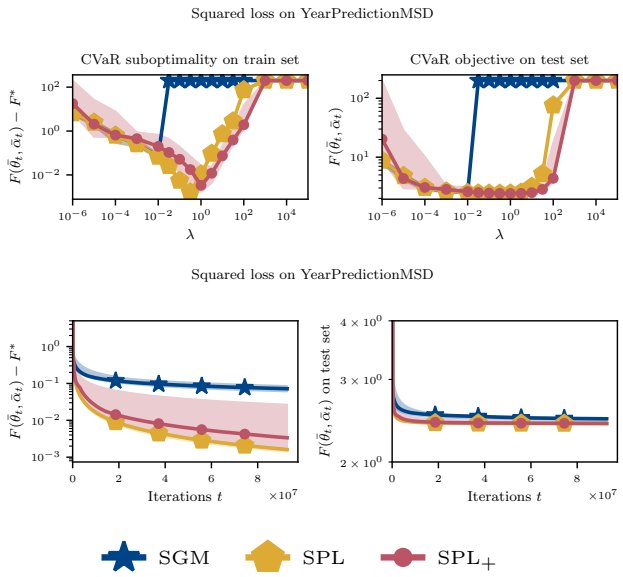

Figure 4: Sensitivity and convergence plots on the `YearPredictionMSD` linear regression task (overdetermined).

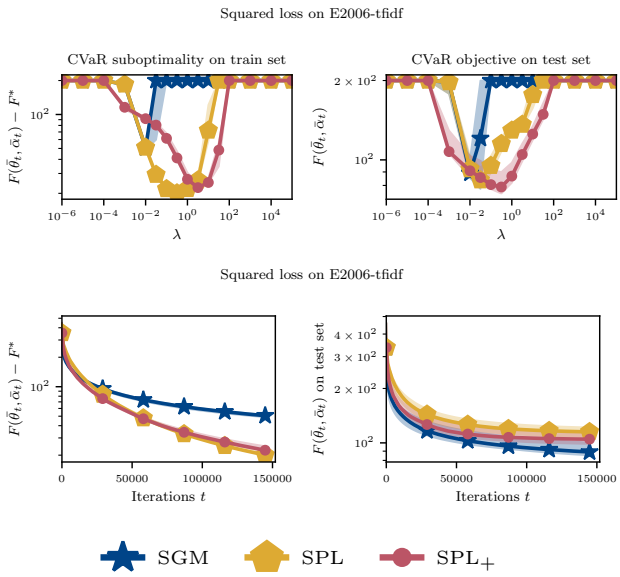

Figure 5: Sensitivity and convergence plots on the `E2006-tfidf` linear regression task (underdetermined).

plots are based on the best λ at the end of training for each method.

For the least squares problem in Figure 4 and Figure 5, we again see that both SPL and SPL+ can tolerate a much larger range of step sizes. The best λ is attained at or near $\lambda = 1$ for SPL+, which, although performs slightly worse than SPL with the best selected λ, allows us to consistently choose $\lambda = 1$ as a default. For the logistic regression problem in Figure 7 and Figure 6, SPL and SPL+ are again similar or better than SGM, although $\lambda = 1$ is no longer close to optimal for SPL and SPL+.

7. Conclusion and future work

Our numerical evidence suggests that for the CVaR minimization problem, while both SGM and SPL can be tuned to achieve similar performance, SPL+ is often the most tolerant to misspecified step sizes. To further speed up SPL+ and make it more competitive over SGM, in future work we will consider using non-uniform sampling to bias towards training examples with higher losses (as in Curi et al. (2020); Sagawa et al. (2020)).

Efficient CVaR minimization with a stochastic algorithm opens up the possibility for new applications in machine learning. For instance, we could consider models that trade-off between low average risk and heavy tails by adding the CVaR objective as a regularizer:

$$\min_{\theta \in \mathbb{R}^d} R_{\mathrm{ERM}}(\theta) + \rho R_{\mathrm{CVaR}_\beta}(\theta)$$

where $\rho > 0$ is a parameter that captures this trade-off. Controlling this trade-off is important as machine learning models are increasingly deployed in safety-critical applications that call for control over the likelihood of failure. As future work, we also see applications in training neural networks, where CVaR can be used to disincentivize the activations from being saturated too often, and thus help in speeding up training. This would offer an alternative to normalization layers, such as batchnorm or layernorm.

Acknowledgements

We would like to thank Vasileios Charisopoulos and Frederik Künstner for helpful feedback on an earlier draft. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), Grant No. PGSD3-547276-2020. This work was partially done during S. Y. Meng's internship at the Flatiron Institute.

References

Andersson, F., Mausser, H., Rosen, D., and Uryasev, S. Credit risk optimization with conditional value-at-risk criterion. *Mathematical programming*, 89(2):273–291, 2001.

Artzner, P., Delbaen, F., Eber, J.-M., and Heath, D. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999.

Asi, H. and Duchi, J. C. The importance of better models

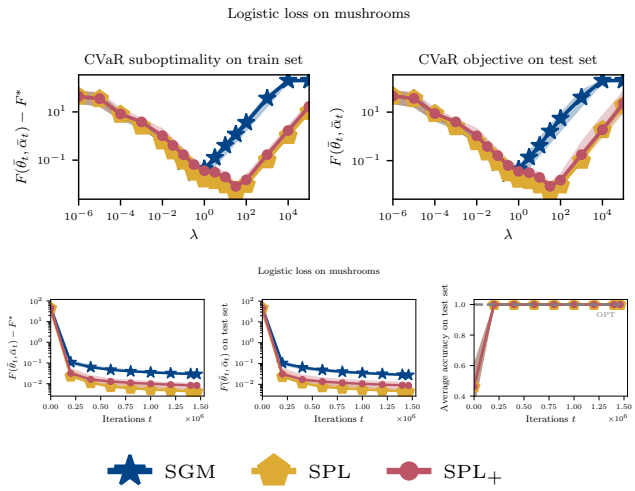

Figure 6: Sensitivity and convergence plots on the `mushrooms` binary classification task. The grey dashed line is the average accuracy on the test set achieved by θ^*.

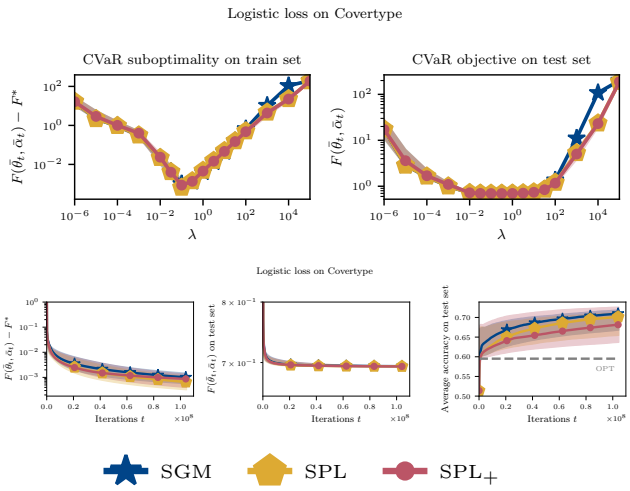

Figure 7: Sensitivity and convergence plots on the `Covertype` binary classification task. The grey dashed line is the average accuracy on the test set achieved by θ^*. Note that the reported accuracy is averaged across the entire training set, but since SPL and SPL+ reached a lower CVaR objective (rather than the average loss objective), it is reasonable that its average accuracy is lower. Furthermore, the optimal accuracy on the test set may seem surprisingly low. The justification behind this is that the objective being minimized is the CVaR objective, while the metric being plotted is the average test accuracy. The CVaR objective puts more emphasis on the top $1-\beta$ fraction of the examples, and so when the dataset is not linearly separable, the average accuracy can be poor. This has also been noted previously in Curi et al. (2020). Unfortunately, computing the accuracy of the examples with the top $1-\beta$ fraction of the losses does not necessarily give us more insight into the accuracy metric. This is due to the possibility that there are only a few outliers and the classifier found by minimizing the CVaR is still getting the majority of the examples wrong, just with lower losses.

in stochastic optimization. *Proceedings of the National Academy of Sciences*, 116(46):22924–22930, 2019a.

Asi, H. and Duchi, J. C. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019b.

Beck, A. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, 2017.

Burke, J. V. and Ferris, M. C. A gauss—newton method for convex composite optimization. *Mathematical Programming*, 71(2):179–194, 1995.

Cardoso, A. R. and Xu, H. Risk-Averse Stochastic Convex Bandit. In *The 22nd International Conference on Artificial Intelligence and Statistics*, volume 89, pp. 39–47, 2019.

Chang, C.-C. and Lin, C.-J. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

Chow, Y. and Ghavamzadeh, M. Algorithms for CVaR Optimization in MDPs. In *Advances in Neural Information Processing Systems 27*, pp. 3509–3517, 2014.

Chow, Y., Tamar, A., Mannor, S., and Pavone, M. Risk-Sensitive and Robust Decision-Making: a CVaR Optimization Approach. In *Advances in Neural Information Processing Systems 28*, pp. 1522–1530, 2015.

Chow, Y., Ghavamzadeh, M., Janson, L., and Pavone, M. Risk-Constrained Reinforcement Learning with Percentile Risk Criteria. *Journal of Machine Learning Research*, 18:167:1–167:51, 2017.

Curi, S., Levy, K. Y., Jegelka, S., and Krause, A. Adaptive sampling for stochastic risk-averse learning. In *Advances in Neural Information Processing Systems 33*, 2020.

Davis, D. and Drusvyatskiy, D. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimimization*, 29(1):207–239, 2019.

Duchi, J. C. and Namkoong, H. Learning models with uniform performance via distributionally robust optimization. *arXiv:1810.08750*, 2018.

Duchi, J. C. and Ruan, F. Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimimization*, 28(4):3229–3259, 2018.

Embrechts, P., Resnick, S. I., and Samorodnitsky, G. Extreme value theory as a risk management tool. *North American Actuarial Journal*, 3(2):30–41, 1999.

Embrechts, P., Klüppelberg, C., and Mikosch, T. *Modelling extremal events: for insurance and finance*, volume 33. Springer Science & Business Media, 2013.

Gotoh, J.-y. and Takeda, A. CVaR minimizations in support vector machines. *Financial Signal Processing and Machine Learning*, pp. 233–265, 2016.

Gower, R. M., Blondel, M., Gazagnadou, N., and Pedregosa, F. Cutting some slack for sgd with adaptive polyak stepsizes, 2022.

Holland, M. and Haress, E. M. Learning with risk-averse feedback under potentially heavy tails. In *International Conference on Artificial Intelligence and Statistics*, pp. 892–900. PMLR, 2021.

Krokhmal, P., Palmquist, J., and Uryasev, S. Portfolio optimization with conditional value-at-risk objective and constraints. *Journal of Risk*, 4:43–68, 2002.

Laguel, Y., Pillutla, K., Malick, J., and Harchaoui, Z. Superquantiles at work: Machine learning applications and efficient subgradient computation. *Set-Valued and Variational Analysis*, 29(4):967–996, 2021a.

Laguel, Y., Pillutla, K., Malick, J., and Harchaoui, Z. A superquantile approach to federated learning with heterogeneous devices. In *55th Annual Conference on Information Sciences and Systems, CISS*, pp. 1–6. IEEE, 2021b.

Lewis, A. S. and Wright, S. J. A proximal method for composite minimization. *Mathematical Programming*, 158:501–546, 2016.

Maehara, T. Risk averse submodular utility maximization. *Operations Research Letters*, 43(5):526–529, 2015.

Mansini, R., Ogryczak, W., and Speranza, M. G. Conditional value at risk and related linear programming models for portfolio optimization. *Annals of Operations Research*, 152(1):227–256, 2007.

Ohsaka, N. and Yoshida, Y. Portfolio optimization for influence spread. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 977–985, 2017.

Rockafellar, R. T. and Uryasev, S. Optimization of Conditional Value-at-Risk. *Journal of Risk*, 2:21–42, 2000.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Sani, A., Lazaric, A., and Munos, R. Risk-Aversion in Multi-armed Bandits. In *Advances in Neural Information Processing Systems 25*, pp. 3284–3292, 2012.

Shalev-Shwartz, S. and Wexler, Y. Minimizing the Maximal Loss: How and Why. In *Proceedings of the 33nd International Conference on Machine Learning*, volume 48 of *JMLR Workshop and Conference Proceedings*, pp. 793–801. JMLR.org, 2016.

Soma, T. and Yoshida, Y. Statistical learning with conditional value at risk. *arXiv:2002.05826*, 2020.

Takeda, A. and Sugiyama, M. ν-support vector machine as conditional value-at-risk minimization. In *Proceedings of the Twenty-Fifth International Conference on Machine Learning*, volume 307, pp. 1056–1063, 2008.

Wilder, B. Risk-Sensitive Submodular Optimization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 6451–6458, 2018.

Williamson, R. C. and Menon, A. K. Fairness risk measures. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6786–6797, 2019.

A. SPL+ derivation for CVaR minimization

Before deriving the updates, we first introduce the following lemma based on the truncated model from Asi & Duchi (2019b).

Lemma 2 (Truncated model)**.** *Consider the problem*

$$x_{t+1} = \arg\min_{x \in \mathbb{R}^n} \; \max\left\{c + \langle a,\, x - x_t\rangle,\, 0\right\} + \frac{1}{2\lambda}\left\|x - x_t\right\|^2.$$

for some scalar c and vector $a \in \mathbb{R}^n$. The solution can be written in closed form as

$$x_{t+1} = x_t - \min\left\{\lambda,\, \frac{\max\{c,\, 0\}}{\|a\|^2}\right\} a$$

Proof. Note that x_{t+1} is the proximal point of the function

$$f(x) = h(\langle a, x\rangle + b), \quad \text{with} \;\; h(z) = \max\left\{z,\, 0\right\},\; b = c - \langle a, x_t\rangle.$$

centered at $x \equiv x_t$. Using Beck (2017, Theorem 6.15), we have

$$\begin{aligned}\operatorname{prox}_{\lambda f}(x) &= x + \frac{a}{\|a\|^2}\left(\operatorname{prox}_{\lambda\|a\|^2 h}(\langle a, x\rangle + b) - (\langle a, x\rangle + b)\right) \\ &= x_t + \frac{a}{\|a\|^2}\left(\operatorname{prox}_{\lambda\|a\|^2 \max\{\cdot,\, 0\}}(c) - c\right)\end{aligned} \tag{28}$$

In turn, the max function is the support function of the interval $[0, 1]$. By Beck (2017, Theorem 6.46), it follows that

$$\operatorname{prox}_{\lambda\|a\|^2 \max\{\cdot,\, 0\}}(c) = c - \lambda\|a\|^2 \operatorname{proj}_{[0,1]}\left(\frac{c}{\lambda\|a\|^2}\right). \tag{29}$$

Plugging (29) into (28), we obtain

$$\begin{aligned}\operatorname{prox}_{\lambda f}(x_t) &= x_t - \frac{a}{\|a\|^2} \cdot \lambda\|a\|^2 \operatorname{proj}_{[0,1]}\left(\frac{c}{\lambda\|a\|^2}\right) \\ &= x_t - \lambda a \cdot \operatorname{proj}_{[0,1]}\left(\frac{c}{\lambda\|a\|^2}\right).\end{aligned}$$

Writing $\operatorname{proj}_{[0,1]}(v) = \min\left\{\max\left\{v,\, 0\right\},\, 1\right\}$ yields the result. □

Lemma 1 (Closed form updates of SPL+)**.** *The closed form solution to* (13) *is given by the updates*

$$\theta_{t+1} = \theta_t - \lambda_{\theta,t} \min\left\{\frac{1}{1-\beta},\, \gamma_t\right\} \nabla\ell(\theta_t; z), \tag{17}$$

$$\alpha_{t+1} = \alpha_t - \lambda_{\alpha,t} + \lambda_{\alpha,t} \min\left\{\frac{1}{1-\beta},\, \gamma_t\right\}, \tag{18}$$

$$\textit{where} \;\; \gamma_t = \frac{\max\left\{\ell(\theta_t; z) - \alpha_t + \lambda_{\alpha,t},\, 0\right\}}{\lambda_{\theta,t}\left\|\nabla\ell(\theta_t; z)\right\|^2 + \lambda_{\alpha,t}}. \tag{19}$$

Proof. We now derive the the SPL+ updates. Recall that for the CVaR objective, using the model m_t^{SPL} in (13), the stochastic model-based approach solves the following problem in Equation 13 at each iteration, that is

$$\arg\min_{\theta,\alpha} \; \alpha + \frac{1}{1-\beta}\max\left\{\ell(\theta_t; z) + \langle v_t,\, \theta - \theta_t\rangle - \alpha, 0\right\} + \frac{1}{2\lambda_\theta}\left\|\theta - \theta_t\right\|^2 + \frac{1}{2\lambda_\alpha}(\alpha - \alpha_t)^2 \tag{30}$$

where $v_t \in \partial \ell(\theta_t; z)$, and we have temporarily dropped the time-dependence on $\lambda_{\alpha,t}$ and $\lambda_{\theta,t}$. To arrive at the closed form solution, we will re-write (30) to fit the format of Lemma 2, and then apply the lemma. To this end, we combine the α in front with its regularization term,

$$\begin{aligned}
\alpha + \frac{1}{2\lambda_\alpha}(\alpha - \alpha_t)^2 &= \alpha + \frac{1}{2\lambda_\alpha}(\alpha^2 - 2\alpha\alpha_t + (\alpha_t)^2) \\
&= \frac{1}{2\lambda_\alpha}((\alpha - \alpha_t)^2 + 2\lambda_\alpha\alpha) \\
&= \frac{1}{2\lambda_\alpha}((\alpha - \alpha_t)^2 + 2\lambda_\alpha\alpha - 2\lambda_\alpha\alpha_t + \lambda_\alpha^2) + \frac{1}{2\lambda_\alpha}(2\lambda_\alpha\alpha_t - \lambda_\alpha^2) \\
&= \frac{1}{2\lambda_\alpha}((\alpha - \alpha_t)^2 + 2\lambda_\alpha(\alpha - \alpha_t) + \lambda_\alpha^2) + \text{Const.} \\
&= \frac{1}{2\lambda_\alpha}(\alpha + \lambda_\alpha - \alpha_t)^2 + \text{Const.}
\end{aligned}$$

We now combine it with the regularization on θ

$$\begin{aligned}
\underbrace{\frac{1}{2\lambda_\theta}\|\theta - \theta_t\|^2 + \alpha + \frac{1}{2\lambda_\alpha}(\alpha - \alpha_t)^2}_{(*)} &= \frac{1}{2\lambda_\theta}\left(\|\theta - \theta_t\|^2 + \frac{\lambda_\theta}{\lambda_\alpha}(\alpha - \alpha_t + \lambda_\alpha)^2\right) + \text{Const.} \\
&= \frac{1}{2\lambda_\theta}\left(\|\theta - \theta_t\|^2 + \left(\sqrt{\frac{\lambda_\theta}{\lambda_\alpha}}(\alpha - \alpha_t) + \sqrt{\lambda_\theta\lambda_\alpha}\right)^2\right) + \text{Const.}
\end{aligned}$$

Now we define a rescaled variable α and constant α_t as

$$\hat{\alpha} = \sqrt{\frac{\lambda_\theta}{\lambda_\alpha}}\alpha \qquad \text{and} \qquad \hat{\alpha}_t = \sqrt{\frac{\lambda_\theta}{\lambda_\alpha}}\alpha_t - \sqrt{\lambda_\theta\lambda_\alpha} \tag{31}$$

to arrive at

$$(*) = \frac{1}{2\lambda_\theta}\left(\|\theta - \theta_t\|^2 + (\hat{\alpha} - \hat{\alpha}_t)^2\right) + \text{Const.}$$

As a side note: to see that the units argument is appropriate, observe that $\hat{\alpha}$ now has units(θ) since λ_θ has units inversely proportional to λ_α. This lets us concatenate α with θ for form a new variable vector $x \in \mathbb{R}^{d+1}$ to have the same units overall. Now define

$$x = \begin{pmatrix} \theta \\ \hat{\alpha} \end{pmatrix} \qquad \text{and} \qquad x_t = \begin{pmatrix} \theta_t \\ \hat{\alpha}_t \end{pmatrix}. \tag{32}$$

The linearization inside $\max\{\cdot, 0\}$ in (30) can be written as

$$\begin{aligned}
\ell(\theta_t; z) + \langle \nabla\ell(\theta_t; z),\, \theta - \theta_t\rangle - \alpha &= \ell(\theta_t; z) + \langle \nabla\ell(\theta_t; z),\, \theta - \theta_t\rangle - \sqrt{\frac{\lambda_\theta}{\lambda_\alpha}}\sqrt{\frac{\lambda_\alpha}{\lambda_\theta}}\alpha \\
&= \ell(\theta_t; z) + \langle \nabla\ell(\theta_t; z),\, \theta - \theta_t\rangle - \sqrt{\frac{\lambda_\alpha}{\lambda_\theta}}\hat{\alpha} + \sqrt{\frac{\lambda_\alpha}{\lambda_\theta}}\hat{\alpha}_t - \sqrt{\frac{\lambda_\alpha}{\lambda_\theta}}\hat{\alpha}_t \\
&= \ell(\theta_t; z) - \sqrt{\frac{\lambda_\alpha}{\lambda_\theta}}\hat{\alpha}_t + \begin{pmatrix} \nabla\ell(\theta_t; z) & -\sqrt{\frac{\lambda_\alpha}{\lambda_\theta}} \end{pmatrix}\begin{pmatrix} \theta - \theta_t \\ \hat{\alpha} - \hat{\alpha}_t \end{pmatrix},
\end{aligned} \tag{33}$$

and so minimizing the model m_t is then equivalent to minimizing the following model $\hat{m}_t$

$$\min_{x \in \mathbb{R}^{d+1}} \max\{c + \langle a,\, x - x_t\rangle,\, 0\} + \frac{1}{2\lambda_\theta}\|x - x_t\|^2$$

up to constants, where

$$c = \frac{1}{1-\beta}\left(\ell(\theta_t; z) - \sqrt{\frac{\lambda_\alpha}{\lambda_\theta}}\hat{\alpha}_t\right) \qquad \text{and} \qquad a = \frac{1}{1-\beta}\begin{pmatrix}\nabla\ell(\theta_t; z) \\ -\sqrt{\frac{\lambda_\alpha}{\lambda_\theta}}.\end{pmatrix}.$$

From Lemma 2, the update is given by

$$x^* = x_t - \eta \cdot a$$

where step size is given by

$$\eta \coloneqq \min\left\{\lambda_\theta,\ \frac{\max\{c,\ 0\}}{\|a\|^2}\right\} \tag{34}$$

Plugging in a, c into η gives

$$\theta_{t+1} = \theta_t - \min\left\{\lambda_\theta,\ \frac{1}{1-\beta}\frac{\max\left\{\ell(\theta_t; z) - \sqrt{\frac{\lambda_\alpha}{\lambda_\theta}}\hat{\alpha}_t,\ 0\right\}}{\frac{1}{(1-\beta)^2}(\|\nabla\ell(\theta_t; z)\|^2 + \frac{\lambda_\alpha}{\lambda_\theta})}\right\}\frac{\nabla\ell(\theta_t; z)}{1-\beta}$$

$$\hat{\alpha}_{t+1} = \hat{\alpha}_t + \frac{1}{1-\beta}\sqrt{\frac{\lambda_\alpha}{\lambda_\theta}}\min\left\{\lambda_\theta,\ \frac{1}{1-\beta}\frac{\max\left\{\ell(\theta_t; z) - \sqrt{\frac{\lambda_\alpha}{\lambda_\theta}}\hat{\alpha}_t,\ 0\right\}}{\frac{1}{(1-\beta)^2}(\|\nabla\ell(\theta_t; z)\|^2 + \frac{\lambda_\alpha}{\lambda_\theta})}\right\}.$$

Finally, substituting back using (31), that is $\hat{\alpha}_{t+1} = \sqrt{\frac{\lambda_\theta}{\lambda_\alpha}}\alpha_{t+1}$ and $\hat{\alpha}_t = \sqrt{\frac{\lambda_\theta}{\lambda_\alpha}}\alpha_t - \sqrt{\lambda_\theta\lambda_\alpha}$ and simplifying gives (17) and (18). □

Lemma 3. *Each SPL+ update in Algorithm 2 is equivalent to the updates given by Equation 18 and Equation 17.*

Proof. We can enumerate all the cases:

1. If $c < 0$, which implies checking for

$$\ell(\theta_t; z) < \sqrt{\frac{\lambda_\alpha}{\lambda_\theta}}\hat{\alpha}_t = \sqrt{\frac{\lambda_\alpha}{\lambda_\theta}}\left(\sqrt{\frac{\lambda_\theta}{\lambda_\alpha}}\alpha_t - \sqrt{\lambda_\theta\lambda_\alpha}\right) = \alpha_t - \lambda_\alpha$$

   then from (34) $\eta = 0$, and the updates are

$$\theta_{t+1} = \theta^* = \theta_t$$
$$\alpha_{t+1} = \hat{\alpha}^* = \hat{\alpha}_t$$

   Multiplying the second equation by $\sqrt{\frac{\lambda_\alpha}{\lambda_\theta}}$ on both sides, we get

$$\sqrt{\frac{\lambda_\alpha}{\lambda_\theta}}\sqrt{\frac{\lambda_\theta}{\lambda_\alpha}}\alpha^* = \sqrt{\frac{\lambda_\alpha}{\lambda_\theta}}\left(\sqrt{\frac{\lambda_\theta}{\lambda_\alpha}}\alpha_t - \sqrt{\lambda_\theta\lambda_\alpha}\right)$$
$$\alpha_{t+1} = \alpha_t - \lambda_\alpha.$$

2. If $c > \lambda_\theta \|a\|^2 \quad (> 0)$, which implies checking for the condition

$$\frac{1}{1-\beta}\left(\ell(\theta_t; z) - \sqrt{\frac{\lambda_\alpha}{\lambda_\theta}}\hat{\alpha}_t\right) > \frac{1}{(1-\beta)^2}\left(\lambda_\theta\|\nabla\ell(\theta_t; z)\|^2 + \lambda_\alpha\right)$$
$$\ell(\theta_t; z) - \alpha_t\lambda_\alpha > \frac{1}{1-\beta}\left(\lambda_\theta\|\nabla\ell(\theta_t; z)\|^2 + \lambda_\alpha\right).$$

Then $\eta = \lambda_\theta$, and the updates reduce to

$$\begin{aligned}\theta_{t+1} &= \theta_t - \lambda_\theta \frac{1}{1-\beta}\nabla \ell(\theta_t; z) \\ \alpha_{t+1} = \hat{\alpha}^* &= \hat{\alpha}_t - \lambda_\theta \frac{1}{1-\beta}\left(-\sqrt{\frac{\lambda_\alpha}{\lambda_\theta}}\right) \\ &= \alpha_t - \lambda_\alpha + \frac{1}{1-\beta}\lambda_\alpha.\end{aligned}$$

3. Otherwise it must be the case that $0 < \frac{c}{\|a\|^2} < \lambda_\theta$, so $\eta = \frac{c}{\|a\|^2}$, and the updates are given by

$$\begin{aligned}\begin{pmatrix}\theta_{t+1}\\ \alpha_{t+1}\end{pmatrix} = \begin{pmatrix}\theta^*\\ \hat{\alpha}^*\end{pmatrix} &= \begin{pmatrix}\theta_t\\ \hat{\alpha}_t\end{pmatrix} - \frac{c}{\|a\|^2}\cdot a \\ &= \begin{pmatrix}\theta_t\\ \hat{\alpha}_t\end{pmatrix} - \frac{\ell(\theta_t; z) - \sqrt{\frac{\lambda_\alpha}{\lambda_\theta}}\hat{\alpha}_t}{\|\nabla \ell(\theta_t; z)\|^2 + \frac{\lambda_\alpha}{\lambda_\theta}}\cdot \begin{pmatrix}\nabla \ell(\theta_t; z)\\ -\sqrt{\frac{\lambda_\alpha}{\lambda_\theta}}\end{pmatrix} \\ &= \begin{pmatrix}\theta_t\\ \hat{\alpha}_t\end{pmatrix} - \underbrace{\frac{\ell(\theta_t; z) - \alpha_t + \lambda_\alpha}{\lambda_\theta \|\nabla \ell(\theta_t; z)\|^2 + \lambda_\alpha}}_{=:\nu}\lambda_\theta \cdot \begin{pmatrix}\nabla \ell(\theta_t; z)\\ -\sqrt{\frac{\lambda_\alpha}{\lambda_\theta}}\end{pmatrix}\end{aligned}$$

Converting $\hat{\alpha}_t$ to α_t and $\hat{\alpha}^*$ to α^*, we get that the updates are

$$\begin{aligned}\theta_{t+1} &= \theta_t - \lambda_\theta \nu \nabla \ell(\theta_t; z) \\ \alpha_{t+1} &= \alpha_t - \lambda_\alpha + \lambda_\alpha \nu.\end{aligned}$$

Note that the regularization parameters λ_θ and λ_α can both be written in a time-dependent form as $\lambda_{\theta,t}$ and $\lambda_{\alpha,t}$. This concludes our derivation for the updates of SPL+ given in Algorithm 2. As a comparison, we also include the closed-form updates for SGM applied to CVaR minimization in Algorithm 1.

□

B. Proof of Theorem 5.2

Theorem 5.2 (Convergence rates of SGM and SPL+)**.** *Suppose Assumption 5.1 holds. Let $x^* = (\theta^*, \alpha^*)^\intercal$ be a minimizer of $F_\beta(\theta, \alpha)$, and $x_0 \in \mathbb{R}^d$ an arbitrary initialization. Let $(x_t)_{t=0}^T$ be the iterates given by SGM or SPL+, and $\bar{x}_T = \frac{1}{T+1}\sum_{t=1}^{T+1} x_t$ be the averaged iterate.*

SGM. *If $\lambda_t = \frac{\lambda}{\sqrt{T+1}}$ then the iterates (x_t) given by SGM in (7) satisfy*

$$\begin{aligned}&\mathbb{E}\left[F_\beta(\bar{x}_T) - F_\beta(x^*)\right]\\ &\quad\le \frac{1}{2}\frac{\|\theta_0 - \theta^*\|^2}{\lambda\sqrt{T+1}} + \frac{1}{2}\frac{(\alpha_0 - \alpha^*)^2}{\lambda\sqrt{T+1}} + \frac{\lambda \mathsf{L}_{\mathrm{SGM}}^2}{\sqrt{T+1}},\end{aligned} \tag{21}$$

where

$$\mathsf{L}_{\mathrm{SGM}}^2 = \mathbb{E}_z\left[\frac{M(z)^2+1}{(1-\beta)^2} + 1\right] \tag{22}$$

SPL+. *If $\lambda_{\alpha,t} = \frac{\lambda_\alpha}{\sqrt{T+1}}$ and $\lambda_{\theta,t} = \frac{\lambda_\theta}{\sqrt{T+1}}$, then the iterates (x_t) given by SPL+ given in Lemma 1 satisfy*

$$\begin{aligned}&\mathbb{E}\left[F_\beta(\bar{x}_T) - F_\beta(x^*)\right]\\ &\quad\le \frac{1}{2}\frac{\|\theta_0 - \theta^*\|^2}{\lambda_\theta\sqrt{T+1}} + \frac{1}{2}\frac{(\alpha_0 - \alpha^*)^2}{\lambda_\alpha\sqrt{T+1}} + \frac{\lambda_\alpha \mathsf{L}_{\mathrm{SPL}}^2}{\sqrt{T+1}},\end{aligned} \tag{23}$$

where

$$\mathsf{L}_{\mathrm{SPL+}}^2 = \mathbb{E}_z\left[\frac{\frac{\lambda_\theta}{\lambda_\alpha}M(z)^2+1}{(1-\beta)^2}\right]. \tag{24}$$

Proof. For our proof, we Recall Equation 30 (restated here)

$$\arg\min_{\theta,\alpha}\ \alpha + \frac{1}{1-\beta}\max\left\{\ell(\theta_t; z) + \langle v_t,\, \theta - \theta_t\rangle - \alpha, 0\right\} + \frac{1}{2\lambda_\theta}\|\theta - \theta_t\|^2 + \frac{1}{2\lambda_\alpha}(\alpha - \alpha_t)^2$$

is the subproblem we solve to obtain the updates with separate regularization. Again, we have temporarily dropped the time-dependency on $\lambda_{\alpha,t}$ and $\lambda_{\theta,t}$. The $\arg\min$ is the same if we scale the entire expression by $\sqrt{\frac{\lambda_\theta}{\lambda_\alpha}}$:

$$\begin{aligned}\arg\min_{\theta,\alpha}\ &\sqrt{\frac{\lambda_\theta}{\lambda_\alpha}}\alpha + \frac{1}{1-\beta}\max\left\{\sqrt{\frac{\lambda_\theta}{\lambda_\alpha}}\left(\ell(\theta_t; z) + \langle v_t,\, \theta - \theta_t\rangle\right) - \sqrt{\frac{\lambda_\theta}{\lambda_\alpha}}\alpha,\, 0\right\}\\ &+ \frac{1}{2\sqrt{\lambda_\theta\lambda_\alpha}}\|\theta - \theta_t\|^2 + \frac{1}{2\lambda_\alpha}\sqrt{\frac{\lambda_\theta}{\lambda_\alpha}}\frac{\lambda_\alpha}{\lambda_\theta}\left(\sqrt{\frac{\lambda_\theta}{\lambda_\alpha}}\alpha - \sqrt{\frac{\lambda_\theta}{\lambda_\alpha}}\alpha_t\right)^2\end{aligned} \tag{35}$$

Let $\hat{\alpha} \coloneqq \sqrt{\frac{\lambda_\theta}{\lambda_\alpha}}\alpha$ and $\hat{\alpha}_t \coloneqq \sqrt{\frac{\lambda_\theta}{\lambda_\alpha}}\alpha_t$ Note that this is a simpler definition of $\hat{\alpha}_t$ than what we used in the derivation of the updates, since we no longer have to absorb the leading α into the regularization. The subproblem (35) can be solved in term of the variables θ and $\hat{\alpha}$, and the scaled linearization

$$\arg\min_{\theta,\hat{\alpha}}\ \hat{\alpha} + \frac{1}{1-\beta}\max\left\{\left(\hat{\ell}(\theta_t; z) + \langle \hat{v}_t,\, \theta - \theta_t\rangle\right) - \hat{\alpha},\, 0\right\} + \frac{1}{2\sqrt{\lambda_\theta\lambda_\alpha}}\left(\|\theta - \theta_t\|^2 + (\hat{\alpha} - \hat{\alpha}_t)^2\right) \tag{36}$$

where $\hat{\ell}(\theta_t; z) \coloneqq \sqrt{\frac{\lambda_\theta}{\lambda_\alpha}}\ell(\theta_t; z)$, and its scaled subgradient is $\hat{v}_t \coloneqq \sqrt{\frac{\lambda_\theta}{\lambda_\alpha}}v_t$. Now define the scaled CVaR objective to be

$$\hat{F}_\beta(\theta, \hat{\alpha}) = \hat{\alpha} + \frac{1}{1-\beta}\mathbb{E}_{z\sim P}\left[\max\left\{\hat{\ell}(\theta; z) - \hat{\alpha},\, 0\right\}\right] \tag{37}$$

and the updates in the scaled subproblem (36) gives the SPL+ method for solving this scaled CVaR problem. By Lemma 4 we have that the assumptions required to invoke Theorem 4.4 in Davis & Drusvyatskiy (2019) now hold. In particular since $\ell(\theta; z)$ is $M(z)$–Lipschitz we have that $\hat{\ell}(\theta; z)$ is $\sqrt{\frac{\lambda_\theta}{\lambda_\alpha}}M(z)$–Lipschitz. We first consider the convergence of SPL+ in terms of the scaled objectives. Denoting $\Delta = \|x_0 - x^*\|$, $\hat{\lambda} \coloneqq \sqrt{\lambda_\theta \lambda_\alpha}$, $x = (\theta, \hat{\alpha})^\top$, $x^* = (\theta^*, \hat{\alpha}^*)^\top$ a minimizer of $\hat{F}_\beta$. Using a constant step size of $\hat{\lambda}_t = \frac{\hat{\lambda}}{\sqrt{T+1}}$, from Theorem 4.4 in Davis & Drusvyatskiy (2019) the convergence rate is

$$\mathbb{E}\left[\hat{F}_\beta(\bar{x}_T) - \hat{F}_\beta(x^*)\right] \leq \frac{\frac{1}{2}\Delta^2 + \mathsf{L}^2_{\mathrm{SPL}+}\hat{\lambda}^2}{\hat{\lambda}\sqrt{T+1}}. \tag{38}$$

Finally, multiplying (37) through by $\sqrt{\frac{\lambda_\alpha}{\lambda_\theta}}$ we have that

$$\sqrt{\frac{\lambda_\alpha}{\lambda_\theta}}\hat{F}_\beta(\theta, \hat{\alpha}) = \alpha + \frac{1}{1-\beta}\mathbb{E}_{z\sim P}\left[\max\left\{\ell(\theta; z) - \alpha,\, 0\right\}\right] = F_\beta(\theta, \alpha).$$

Furthermore, multiplying (39) through by $\sqrt{\frac{\lambda_\alpha}{\lambda_\theta}}$ and substituting back $\hat{\lambda} \coloneqq \sqrt{\lambda_\theta \lambda_\alpha}$ and

$$\Delta^2 = \|\theta_0 - \theta^*\|^2 + (\hat{\alpha}_0 - \hat{\alpha}^*)^2 = \|\theta_0 - \theta^*\|^2 + \frac{\lambda_\theta}{\lambda_\alpha}(\alpha_0 - \alpha^*)^2$$

gives

$$\mathbb{E}\left[F_\beta(\bar{x}_T) - F_\beta(x^*)\right] \leq \sqrt{\frac{\lambda_\alpha}{\lambda_\theta}}\frac{\frac{1}{2}\Delta^2 + \mathsf{L}^2_{\mathrm{SPL}+}\lambda_\theta\lambda_\alpha}{\sqrt{\lambda_\theta\lambda_\alpha}\sqrt{T+1}} \tag{39}$$

$$= \frac{\frac{1}{2}\|\theta_0 - \theta^*\|^2 + \frac{1}{2}\frac{\lambda_\theta}{\lambda_\alpha}(\alpha_0 - \alpha^*)^2 + \mathsf{L}^2_{\mathrm{SPL}+}\lambda_\theta\lambda_\alpha}{\lambda_\theta\sqrt{T+1}} \tag{40}$$

$$= \frac{1}{2}\frac{\|\theta_0 - \theta^*\|^2}{\lambda_\theta\sqrt{T+1}} + \frac{1}{2}\frac{(\alpha_0 - \alpha^*)^2}{\lambda_\alpha\sqrt{T+1}} + \frac{\mathsf{L}^2_{\mathrm{SPL}+}\lambda_\alpha}{\sqrt{T+1}} \tag{41}$$

which concludes the proof of convergence of SPL+. As for the proof of SGM, it only remains to choose $\lambda_\theta = \lambda_\alpha = \lambda$ □

To apply Theorem 4.4 in Davis & Drusvyatskiy (2019), we must first verify their assumptions (B1)-(B4) hold. We will enumerate these under their following general setup: writing the CVaR objective in Equation 5 as

$$F_\beta(x) = f(x) + r(x), \tag{42}$$

where $r(x) = 0$ for SGM while $r(x) = \hat{\alpha}$ for SPL+. In the SPL+ case, we further write $f(x) = \mathbb{E}_z[h(c(x; z))]$ where $h(\cdot) = \frac{1}{1-\beta}\max\{\cdot, 0\}$ and $c(x; z) = \hat{\ell}(\theta; z) - \hat{\alpha}$. Recall that the stochastic one-sided models used are

$$\text{SGM} \qquad f_t^{\mathrm{SGM}}(x; z) = F_\beta(x_t; z) + \langle g_t,\, x - x_t\rangle \qquad \text{where } g_t \in \partial F_\beta(x_t; z), \quad x = (\theta, \alpha)^\top \tag{43}$$

$$\text{SPL+} \qquad f_t^{\mathrm{SPL}}(x; z) = h(c(x_t; z) + \langle u_t,\, x - x_t\rangle) \qquad \text{where } u_t \in \partial c(x_t; z), \quad x = (\theta, \hat{\alpha})^\top \tag{44}$$

and the update in Equation 11 is equivalent to

$$x_{t+1} = \operatorname*{arg\,min}_{x\in\mathbb{R}^{d+1}} r(x) + f_t(x; z) + \frac{1}{2\lambda_t}\|x - x_t\|^2 \tag{45}$$

The assumptions we need to verify are given in the following Lemma, which are adapted from Davis & Drusvyatskiy (2019).

Lemma 4. *Let $\ell(\theta; z)$ be $M(z)$–Lipschitz and convex. Consider the two alternative definitions for $f_t(x; z)$ given in (43) and (44). We have that the following assumptions hold.*

*(B1) (**Sampling**) It is possible to generate i.i.d. realizations $z_1, z_2, \cdots \sim P$.*

(B2) ***(One-sided accuracy)*** *There is an open set U containing $\operatorname{dom} r$ and a measurable function $(x, y; z) \mapsto g_x(y; z)$, defined on $U \times U \times \Omega$, satisfying*

$$\mathbb{E}_z\left[f_t(x_t; z)\right] = f(x_t) \quad \forall x_t \in U,$$

and

$$\mathbb{E}_z\left[f_t(x; z) - f(x)\right] \leq \frac{\tau}{2}\left\|x_t - x\right\|^2 \quad \forall x_t, x \in U.$$

(B3) ***(Weak-convexity)*** *The function $f_t(x; z) + r(x)$ is η-weakly convex for all $x \in U$, a.e. $z \in \Omega$.*

(B4) ***(Lipschitz property)*** *There exists a measurable function $L : \Omega \to \mathbb{R}_+$ satisfying $\sqrt{\mathbb{E}_z[L(z)^2]} \leq \mathsf{L}$ and such that*

$$f_t(x_t; z) - f_t(x; z) \leq L(z)\left\|x_t - x\right\| \quad \forall x_t, x \in U \text{ and a.e. } z \sim P,$$

where

$$\mathsf{L}_{\mathrm{SGM}}^2 = \mathbb{E}_z\left[\frac{M(z)^2 + 1}{(1-\beta)^2} + 1\right] \qquad \text{for } \textbf{SGM} \text{ where } f_t(x; z) \text{ is (43)}, \tag{46}$$

$$\mathsf{L}_{\mathrm{SPL}+}^2 = \mathbb{E}_z\left[\frac{\frac{\lambda_\theta}{\lambda_\alpha}M(z)^2 + 1}{(1-\beta)^2}\right] \qquad \text{for } \textbf{SPL}+ \text{ where } f_t(x; z) \text{ is (44)}. \tag{47}$$

Proof. Assumption (B1) follows trivially from i.i.d. sampling, while (B2) follows from convexity of $\ell(\cdot; z)$ or $\hat{\ell}(\cdot; z)$, giving us $\tau = 0$. Since $r(x)$ is also convex in both methods and both models are convex, (B3) holds with $\eta = 0$.

To prove item (B4) for SGM, where $f^t = f_t^{\mathrm{SGM}}$ is given in (43), first note that from (6) for $g_t \in \partial F_\beta(x_t; z)$ and $u_t \in \partial\ell(\theta_t; z)$ we have that

$$\begin{aligned}
\left\|g_t\right\|^2 &= \mathbb{1}\left\{\ell(\theta_t; z) - \alpha_t \geq 0\right\}\frac{\left\|u_t\right\|^2}{(1-\beta)^2} + \left(1 - \frac{\mathbb{1}\left\{\ell(\theta_t; z) - \alpha_t \geq 0\right\}}{(1-\beta)}\right)^2 \\
&\leq \mathbb{1}\left\{\ell(\theta_t; z) - \alpha_t \geq 0\right\}\frac{M(z)^2}{(1-\beta)^2} + 1 - 2\frac{\mathbb{1}\left\{\ell(\theta_t; z) - \alpha_t \geq 0\right\}}{(1-\beta)} + \frac{\left(\mathbb{1}\left\{\ell(\theta_t; z) - \alpha_t \geq 0\right\}\right)^2}{(1-\beta)^2} \\
&\leq \frac{M(z)^2}{(1-\beta)^2} + 1 + \frac{1}{(1-\beta)^2},
\end{aligned} \tag{48}$$

where in the first inequality we used that $\ell(\cdot; z)$ is $M(z)$–Lipschitz to bound $\|u_t\| \leq M(z)$, and in the second inequality we used that the indicator function $\mathbb{1}\left\{\ell(\theta_t; z) - \alpha_t \geq 0\right\}$ is positive and upper bounded by 1. Consequently,

$$\left\|g_t\right\| \leq \sqrt{1 + \frac{M(z)^2 + 1}{(1-\beta)^2}}. \tag{49}$$

Thus using the above and that $\max\{\cdot, 0\}$ is 1-Lipschitz:

$$\begin{aligned}
f_t(x_t; z) - f_t(y; z) &\leq \left\|g_t\right\|\left\|x_t - y\right\| \\
&\leq \underbrace{\sqrt{1 + \frac{M(z)^2 + 1}{(1-\beta)^2}}}_{=: L(z)}\left\|x_t - y\right\|. \qquad \text{(By (49))}
\end{aligned}$$

This gives us $\mathsf{L}_{\mathrm{SGM}}^2 = \mathbb{E}_z[L(z)^2] = \mathbb{E}_z\left[1 + \frac{M(z)^2+1}{(1-\beta)^2}\right]$.

For SPL+ and $f^t = f_t^{\mathrm{SPL}}$ defined in (44) we have that

$$\begin{aligned}
(1-\beta)(f_t(x_t;z) - f_t(y;z)) &= \max\left\{\hat{\ell}(\theta_t;z) - \hat{\alpha}_t,\, 0\right\} - \max\left\{\hat{\ell}(\theta_t;z) - \langle \hat{v}_t,\, \theta - \theta_t\rangle - \hat{\alpha},\, 0\right\} \\
&\le \max\left\{\langle \hat{v}_t,\, \theta - \theta_t\rangle + (\hat{\alpha} - \hat{\alpha}_t),\, 0\right\} \qquad (\max\{a,0\} - \max\{b,0\} \le \max\{a-b,0\}) \\
&= \max\left\{\left\langle \begin{pmatrix}\hat{v}_t \\ 1\end{pmatrix}, \begin{pmatrix}\theta - \theta_t \\ \hat{\alpha} - \hat{\alpha}_t\end{pmatrix}\right\rangle, 0\right\} \\
&\le \left\|\begin{pmatrix}\hat{v}_t \\ 1\end{pmatrix}\right\| \left\|\begin{pmatrix}\theta - \theta_t \\ \hat{\alpha} - \hat{\alpha}_t\end{pmatrix}\right\| \qquad \text{(Cauchy-Schwarz)} \\
&\le \sqrt{1 + \|\hat{v}_t\|^2}\, \|x_t - y\| \\
&\le \sqrt{1 + \frac{\lambda_\theta}{\lambda_\alpha} M(z)^2}\, \|x_t - y\| \qquad \text{(Since } \hat{v}_t \text{ is scaled } v_t\text{)}
\end{aligned}$$

Dividing both sides by $(1-\beta)$ gives us

$$f_t(x_t;z) - f_t(y;z) \le \underbrace{\left(\frac{\sqrt{\frac{\lambda_\theta}{\lambda_\alpha} M(z)^2 + 1}}{1-\beta}\right)}_{=:L(z)} \|x_t - y\|.$$

Taking expectation over z yields

$$\mathsf{L}^2_{\mathrm{SPL}+} = \mathbb{E}_z[L(z)^2] = \mathbb{E}_z\left[\frac{\frac{\lambda_\theta}{\lambda_\alpha} M(z)^2 + 1}{(1-\beta)^2}\right].$$

□

C. Additional experiment results

Figure 8 shows a similar sensitivity analysis to Figure 3 in the main text. Instead of the sensitivity of final suboptimality, here we show the sensitivity of the minimum number of iterations to reach ϵ-suboptimality $\tilde{F}(\theta, \alpha) - \tilde{F}^* \leq \epsilon$.

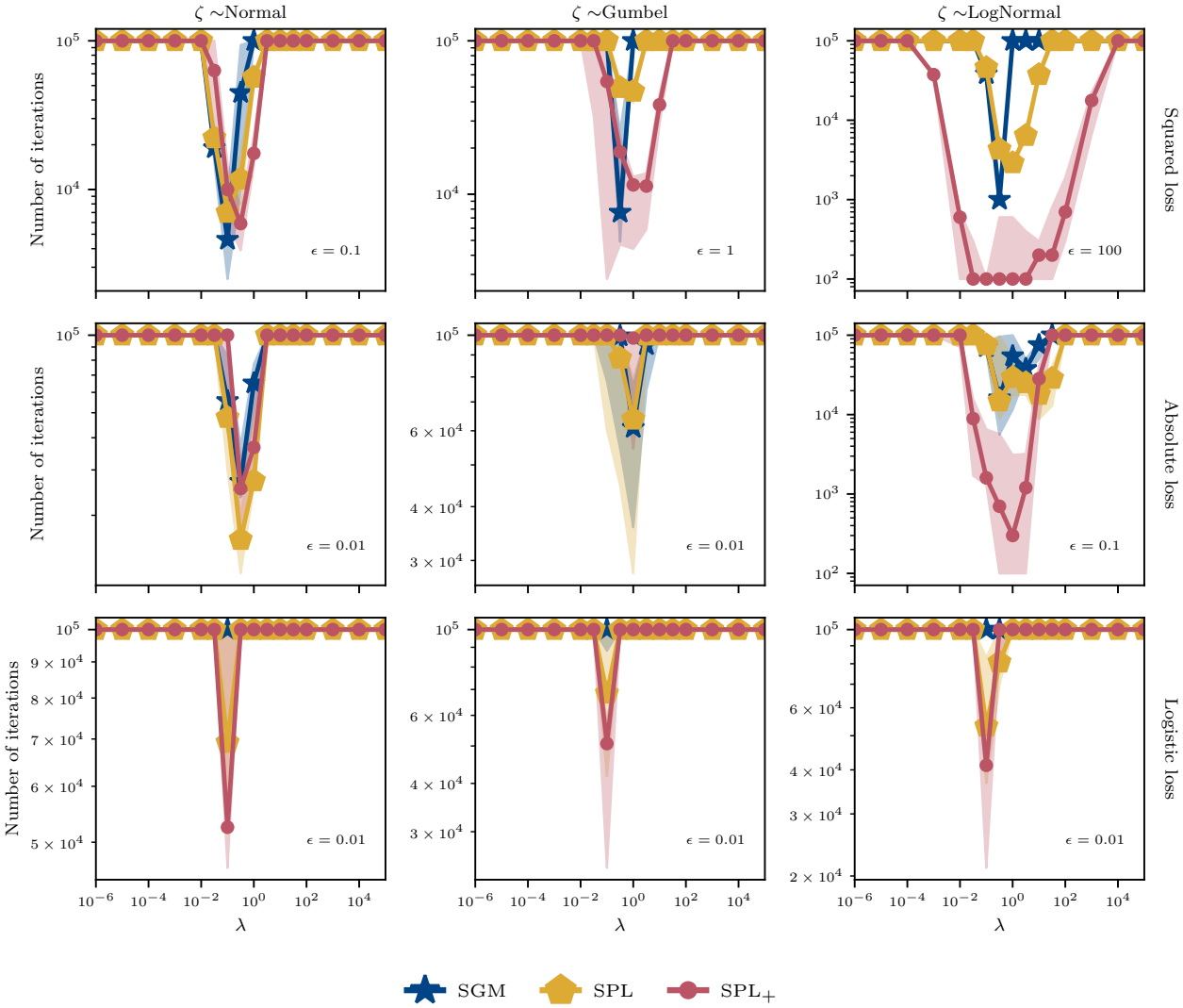

Figure 8: Sensitivity of minimum number of iterations to achieve ϵ suboptimality to step size choices. The first two rows are regression tasks under the ℓ_1 and ℓ_2 losses, while the third row correspond to a binary classification task under the logistic loss. The columns correspond to different noise distributions in the data generation that controls the difficulty of the problem.

D. Relationship to max loss minimization

Lemma 5. *The SPL+ updates in Lemma 1 minimizes the prox-linear model of the Lagrangian of the max loss objective,*

$$\min_{\theta\in\mathbb{R}^d} \; f(\theta) = \max_{i=1,\dots,n} \ell(\theta; z_i)$$

with $\beta = 1 - 1/n$.

Proof. The equivalent slack formulation to the max loss objective is

$$\begin{aligned}
&\min_{s,\theta} \; s \\
&\text{s.t.} \;\; \ell(\theta; z_i) \le s \quad \forall i = 1, \dots, n
\end{aligned}$$

Note that we can add a dummy constraint to have the equivalent problem

$$\begin{aligned}
&\min_{s,w} \; s \\
&\text{s.t.} \;\; \ell(\theta; z_i) \le s \quad \forall i = 1, \dots, n \\
&\qquad 0 \le 0 \\
&\qquad \Updownarrow \\
&\min_{s,w} \; s \\
&\text{s.t.} \;\; \max\{\ell(\theta; z_i) - s,\, 0\} \le 0
\end{aligned}$$

Then the Lagrangian is given by

$$\mathcal{L}(s, \theta, \Gamma) = s + \frac{1}{n}\sum_{i=1}^{n} \Gamma_i \max\{\ell(\theta; z_i) - s,\, 0\} \tag{50}$$

Note that we have included a $1/n$ scaling for each constraint, which is fine because they are positive and so can be absorbed into the Lagrange multipliers. The dual problem is given by

$$\begin{aligned}
&\max_{\Gamma\in\mathbb{R}^n} \; g(\Gamma) = \min_{s,\theta} \mathcal{L}(s, \theta, \Gamma) \\
&\text{s.t.} \;\; \Gamma_i \ge 0 \quad \forall i = 1, \dots, n
\end{aligned}$$

And so given a set of Γ, we need to minimize the Lagrangian over s and θ. We can treat this as the *base objective*, the basis of our stochastic model construction. At each iteration t, we will use the following model

$$m_t(s, \theta) = s + \Gamma_i \max\{\ell(\theta_t; z_i) + \langle \nabla\ell(\theta_t; z_i),\, \theta - \theta_t\rangle - s,\, 0\}$$

And then using the stochastic model-based approach, the updates are given by

$$\theta_{t+1}, s_{t+1} = \arg\min_{s,\theta} \; m_t(s, \theta) + \frac{1}{2\lambda_\theta}\|\theta - \theta_t\|^2 + \frac{1}{2\lambda_s}(s - s_t)^2$$

Observe that this corresponds exactly to our SPL+ updates for the CVaR objective. Specifically, we can recover that by taking $s = \alpha$ and $\Gamma_i = \frac{1}{1-\beta}$. Now let's analyze the KKT conditions to see what Γ should be. Let $u_i^* = \partial \max\{u, 0\}|_{u=f(\theta^*; z_i)-s^*}$

$$\begin{aligned}
0 \in \partial_s \mathcal{L}(s^*, \theta^*, \Gamma^*) = 1 - \frac{1}{n}\sum_{i=1}^{n} \Gamma_i^* u_i^* \iff 1 \in \frac{1}{n}\sum_{i=1}^{n} \Gamma_i^* u_i^* \\
0 \in \partial_\theta \mathcal{L}(s^*, \theta^*, \Gamma^*) = \frac{1}{n}\sum_{i=1}^{n} \Gamma_i^* u_i^* \nabla\ell(\theta^*; z_i) \\
\max\{\ell(\theta^*; z_i) - s^*,\, 0\} \le 0 \qquad \forall i \\
\Gamma_i^* \ge 0 \qquad \forall i \\
\Gamma_i^*(\max\{\ell(\theta^*; z_i) - s^*,\, 0\}) = 0 \qquad \forall i
\end{aligned}$$

First, based on the constraints, we must have $\ell(\theta^*; z_i) \leq s^*$ for all i. Now suppose none of the constraints are tight, i.e. $\ell(\theta^*; z_i) < s^*$ for all i. Then $u_i^* = 0$ for all i so the first KKT condition would collapse. This means that the active constraint set $\mathcal{I} = \{i = 1, \ldots, n : \ell(\theta^*; z_i) = s^*\}$ must be non-empty. We can use this to simplify the first two conditions to

$$1 \in \frac{1}{n} \sum_{i \in \mathcal{I}} \Gamma_i^* [0, 1]$$
$$0 \in \frac{1}{n} \sum_{i \in \mathcal{I}} \Gamma_i^* [0, 1] \nabla \ell(\theta^*; z_i)$$

which holds for $\Gamma_i^* \geq n$ for $i \in \mathcal{I}$. If we let $\Gamma_i^* = n$ for all i, then we will need $\beta = 1 - 1/n$ to recover the CVaR objective, which makes sense in the max loss minimization problem with n training examples, so we can take $P = P_n$ the empirical distribution. □