Bayesian Counterfactual Risk Minimization (Supplemental Material)

Ben London Ted Sandler

A. Proof of Theorem 1

Our proof will leverage the following version of the PAC-Bayesian theorem, due to McAllester (2003).

Lemma 3. *Let $\mathbb{D}$ denote a fixed distribution on an instance space, $\mathcal{Z}$. Let $L : \mathcal{H} \times \mathcal{Z} \to [0,1]$ denote a* loss function*. For a distribution, $\mathbb{Q}$, on the hypothesis space, $\mathcal{H}$, and a dataset, $S \triangleq (z_1, \ldots, z_n) \in \mathcal{Z}^n$, let $R(\mathbb{Q}) \triangleq \mathbb{E}_{h\sim\mathbb{Q}} \mathbb{E}_{z\sim\mathbb{D}}[L(h,z)]$ and $\hat{R}(\mathbb{Q}, S) \triangleq \mathbb{E}_{h\sim\mathbb{Q}} \left[\frac{1}{n}\sum_{i=1}^n L(h, z_i)\right]$ denote the risk and empirical risk, respectively. For any $n \geq 1$, $\delta \in (0,1)$, and fixed prior, $\mathbb{P}$, on $\mathcal{H}$, with probability at least $1-\delta$ over draws of $S \sim \mathbb{D}^n$, the following holds simultaneously for all posteriors, $\mathbb{Q}$, on $\mathcal{H}$:*

$$R(\mathbb{Q}) \leq \hat{R}(\mathbb{Q}, S) + \sqrt{\frac{2\hat{R}(\mathbb{Q}, S)\left(D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{n}{\delta}\right)}{n-1}} + \frac{2\left(D_{\text{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{n}{\delta}\right)}{n-1}.$$

To apply Lemma 3, we need to define an appropriate loss for CRM. It should be expressed as a function of a hypothesis and a single example[1], and bounded in $[0,1]$. Accordingly, we define

$$L_\tau(h, x, a, p, r) \triangleq 1 - \tau\, r\, \frac{\mathbb{1}\{h(x) = a\}}{\max\{p, \tau\}},$$

which satisfies these criteria. Using this loss function, we let

$$R_\tau(\mathbb{Q}) \triangleq \mathop{\mathbb{E}}_{h\sim\mathbb{Q}} \mathop{\mathbb{E}}_{(x,\rho)\sim\mathbb{D}} \mathop{\mathbb{E}}_{a\sim\pi_0(x)} [L_\tau(h, x, a, \pi_0(a\,|\,x), \rho(x,a))]$$

and

$$\hat{R}_\tau(\mathbb{Q}, S) \triangleq \mathop{\mathbb{E}}_{h\sim\mathbb{Q}} \left[\frac{1}{n}\sum_{i=1}^n L_\tau(h, x_i, a_i, p_i, r_i)\right].$$

Importantly, $\hat{R}_\tau(\mathbb{Q}, S)$ is an unbiased estimate of $R_\tau(\mathbb{Q})$,

$$\mathop{\mathbb{E}}_{S\sim(\mathbb{D}\times\pi_0)^n} [\hat{R}_\tau(\mathbb{Q}, S)] = R_\tau(\mathbb{Q}),$$

and a draw of $h \sim \mathbb{Q}$ does not depend on context, so $R_\tau(\mathbb{Q})$ and $\hat{R}_\tau(\mathbb{Q}, S)$ can be expressed as expectations over $h \sim \mathbb{Q}$.[2] Further, via linearity of expectation,

$$\begin{aligned}
R_\tau(\mathbb{Q}) &= 1 - \tau \mathop{\mathbb{E}}_{(x,\rho)\sim\mathbb{D}} \mathop{\mathbb{E}}_{a\sim\pi_0(x)} \left[\rho(x,a) \frac{\mathbb{E}_{h\sim\mathbb{Q}}[\mathbb{1}\{h(x) = a\}]}{\max\{\pi_0(a\,|\,x), \tau\}}\right] \\
&= 1 - \tau \mathop{\mathbb{E}}_{(x,\rho)\sim\mathbb{D}} \mathop{\mathbb{E}}_{a\sim\pi_0(x)} \left[\rho(x,a)\, \frac{\pi_\mathbb{Q}(a\,|\,x)}{\max\{\pi_0(a\,|\,x), \tau\}}\right] \\
&\geq 1 - \tau \mathop{\mathbb{E}}_{(x,\rho)\sim\mathbb{D}} \mathop{\mathbb{E}}_{a\sim\pi_\mathbb{Q}(x)} [\rho(x,a)] \\
&= 1 - \tau\,(1 - R(\pi_\mathbb{Q}))\,,
\end{aligned}$$

[1] This criterion ensures that the (empirical) risk decomposes as a sum of i.i.d. random variables, which is our motivation for using the truncated IPS estimator over the self-normalizing estimator (Swaminathan and Joachims, 2015); the latter does not decompose.

[2] This is why we truncate with $\max\{p_i, \tau\}^{-1}$ instead of $\min\{\pi(a_i\,|\,x_i)/p_i, \tau^{-1}\}$.

and

$$\begin{aligned}\hat{R}_\tau(\mathbb{Q}, S) &= 1 - \frac{\tau}{n}\sum_{i=1}^{n} r_i \, \frac{\mathbb{E}_{h\sim\mathbb{Q}}\left[\mathbb{1}\{h(x_i) = a_i\}\right]}{\max\{p_i, \tau\}} \\ &= 1 - \frac{\tau}{n}\sum_{i=1}^{n} r_i \, \frac{\pi_{\mathbb{Q}}(a_i \,|\, x_i)}{\max\{p_i, \tau\}} \\ &= 1 - \tau\left(1 - \hat{R}_\tau(\pi_{\mathbb{Q}}, S)\right).\end{aligned}$$

Thus,

$$R_\tau(\mathbb{Q}) - \hat{R}_\tau(\mathbb{Q}, S) \geq \tau\big(R(\pi_{\mathbb{Q}}) - \hat{R}_\tau(\pi_{\mathbb{Q}}, S)\big),$$

which means that Lemma 3 can be used to upper-bound $R(\pi_{\mathbb{Q}}) - \hat{R}_\tau(\pi_{\mathbb{Q}}, S)$.

B. Risk Bound for All Truncation Parameters

Since Theorem 1 assumes that the truncation parameter, τ, is fixed *a priori*, we now derive a risk bound that holds for all τ simultaneously. An implication of this bound is that the truncation can be data-dependent.

Theorem 4. *Let $\mathcal{H} \subseteq \{h : \mathcal{X} \to \mathcal{A}\}$ denote a hypothesis space mapping contexts to actions. For any $n \geq 1$, $\delta \in (0, 1)$ and fixed prior, $\mathbb{P}$, on $\mathcal{H}$, with probability at least $1 - \delta$ over draws of $S \sim (\mathbb{D} \times \pi_0)^n$, the following holds simultaneously for all posteriors, $\mathbb{Q}$, on $\mathcal{H}$, and all $\tau \in (0, 1)$:*

$$R(\pi_{\mathbb{Q}}) \leq \hat{R}_\tau(\pi_{\mathbb{Q}}, S) + \sqrt{\frac{4\big(\hat{R}_\tau(\pi_{\mathbb{Q}}, S) - 1 + \frac{2}{\tau}\big)\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2n}{\delta\tau}\right)}{\tau\,(n-1)}} + \frac{4\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2n}{\delta\tau}\right)}{\tau\,(n-1)}.$$

Proof. We construct an infinite sequence of τ values, $(\tau_i \triangleq 2^{-i})_{i=1}^{\infty}$, and δ values, $(\delta_i \triangleq \delta\tau_i)_{i=1}^{\infty}$. For any τ_i, Equation 4 holds with probability at least $1 - \delta_i$. Thus, with probability at least $1 - \sum_{i=0}^{\infty}\delta_i = 1 - \delta$, Equation 4 holds for all τ_i simultaneously.

For a given τ—which may depend on the data—we select $i^\star \triangleq \left\lceil \frac{\ln \tau^{-1}}{\ln 2} \right\rceil$. (Since $\tau \in (0, 1)$, the ceiling function ensures that $i^\star \geq 1$.) Then, we have that $\tau/2 \leq \tau_{i^\star} \leq \tau$; and, since $\max\{p, \tau_{i^\star}\} \leq \max\{p, \tau\}$, we have that $\hat{R}_{\tau_{i^\star}}(\pi, S) \leq \hat{R}_\tau(\pi, S)$. Further, $\delta_{i^\star} \geq \delta\tau/2$. Thus, with probability at least $1 - \delta$,

$$\begin{aligned}R(\pi) &\leq \hat{R}_{\tau_{i^\star}}(\pi, S) + \sqrt{\frac{2\big(\hat{R}_{\tau_{i^\star}}(\pi, S) - 1 + \frac{1}{\tau_{i^\star}}\big)\big(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{n}{\delta_{i^\star}}\big)}{\tau_{i^\star}\,(n-1)}} + \frac{2\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{n}{\delta_{i^\star}}\right)}{\tau_{i^\star}\,(n-1)} \\ &\leq \hat{R}_\tau(\pi, S) + \sqrt{\frac{4\big(\hat{R}_\tau(\pi, S) - 1 + \frac{2}{\tau}\big)\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2n}{\delta\tau}\right)}{\tau\,(n-1)}} + \frac{4\left(D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) + \ln\frac{2n}{\delta\tau}\right)}{\tau\,(n-1)},\end{aligned}$$

which completes the proof. □

C. Proof of Lemma 1

We can ignore the Gumbel distributions, since they are identical. Using the definition of the KL divergence for multivariate Gaussians, and properties of diagonal matrices (since both covariances are diagonal), we have that

$$D_{\mathrm{KL}}(\mathbb{Q}\|\mathbb{P}) = \frac{\|\mu - \mu_0\|^2}{2\sigma_0^2} + \frac{d}{2}\left(\ln\frac{\sigma_0^2}{\sigma^2} + \frac{\sigma^2}{\sigma_0^2} - 1\right).$$

We conclude by noting that $\frac{\sigma^2}{\sigma_0^2} - 1 \leq 0$ for $\sigma^2 \leq \sigma_0^2$.

D. Proof of Lemma 2

We begin with the lower bound. First, let

$$\Phi(w) \triangleq \sum_{a' \in \mathcal{A}} \exp(w \cdot \phi(x, a'))$$

denote a normalizing constant, sometimes referred to as the *partition function*. (Since x is given, our notation ignores the fact that Φ is a function of x.) Using Φ in the definition of ς, and applying Jensen's inequality, we have that

$$\begin{aligned}
\pi_{\mathbb{Q}}(a \,|\, x) &= \mathop{\mathbb{E}}_{w \sim \mathcal{N}(\mu, \sigma^2 \mathbf{I})} \left[\varsigma_w(a \,|\, x)\right] \\
&= \mathop{\mathbb{E}}_{w \sim \mathcal{N}(\mu, \sigma^2 \mathbf{I})} \left[\exp\left(w \cdot \phi(x, a) - \ln \Phi(w)\right)\right] \\
&\geq \exp\left(\mathop{\mathbb{E}}_{w \sim \mathcal{N}(\mu, \sigma^2 \mathbf{I})} \left[w \cdot \phi(x, a) - \ln \Phi(w)\right]\right). \qquad (18)
\end{aligned}$$

We then express the random parameters, $w \sim \mathcal{N}(\mu, \sigma^2 \mathbf{I})$, as the sum of the mean parameters, μ, and a zero-mean Gaussian vector, $g \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$, which yields

$$\begin{aligned}
\mathop{\mathbb{E}}_{w \sim \mathcal{N}(\mu, \sigma^2 \mathbf{I})} \left[w \cdot \phi(x, a) - \ln \Phi(w)\right] &= \mathop{\mathbb{E}}_{g \sim \mathcal{N}(0, \sigma^2 \mathbf{I})} \left[(\mu + g) \cdot \phi(x, a) - \ln \Phi(\mu + g)\right] \\
&= \mu \cdot \phi(x, a) \; - \mathop{\mathbb{E}}_{g \sim \mathcal{N}(0, \sigma^2 \mathbf{I})} \left[\ln \Phi(\mu + g)\right] \\
&= \mu \cdot \phi(x, a) - \ln \Phi(\mu) - \mathop{\mathbb{E}}_{g \sim \mathcal{N}(0, \sigma^2 \mathbf{I})} \left[\ln \left(\frac{\Phi(\mu + g)}{\Phi(\mu)}\right)\right]. \qquad (19)
\end{aligned}$$

The second line follows from the fact that the expected dot product of any vector with a zero-mean Gaussian vector is zero. Applying Jensen's inequality again to the last term, we have

$$- \mathop{\mathbb{E}}_{g \sim \mathcal{N}(0, \sigma^2 \mathbf{I})} \left[\ln \left(\frac{\Phi(\mu + g)}{\Phi(\mu)}\right)\right] \geq - \ln \mathop{\mathbb{E}}_{g \sim \mathcal{N}(0, \sigma^2 \mathbf{I})} \left[\frac{\Phi(\mu + g)}{\Phi(\mu)}\right]. \qquad (20)$$

Observe that

$$\frac{\Phi(\mu + g)}{\Phi(\mu)} = \sum_{a' \in \mathcal{A}} \frac{\exp(\mu \cdot \phi(x, a'))}{\Phi(\mu)} \exp(g \cdot \phi(x, a')) = \mathop{\mathbb{E}}_{a' \sim \varsigma_\mu(x)} \left[\exp(g \cdot \phi(x, a'))\right].$$

Thus, via linearity of expectation,

$$\mathop{\mathbb{E}}_{g \sim \mathcal{N}(0, \sigma^2 \mathbf{I})} \left[\frac{\Phi(\mu + g)}{\Phi(\mu)}\right] = \mathop{\mathbb{E}}_{a' \sim \varsigma_\mu(x)} \mathop{\mathbb{E}}_{g \sim \mathcal{N}(0, \sigma^2 \mathbf{I})} \left[\exp(g \cdot \phi(x, a'))\right]. \qquad (21)$$

The right-hand inner expectation is simply the moment-generating function of a multivariate Gaussian. Combining its closed-form expression,

$$\mathop{\mathbb{E}}_{g \sim \mathcal{N}(0, \sigma^2 \mathbf{I})} \left[\exp(g \cdot \phi(x, a'))\right] = \exp\left(\frac{\sigma^2}{2} \left\|\phi(x, a')\right\|^2\right),$$

with Equation 21, we have

$$\begin{aligned}
- \ln \mathop{\mathbb{E}}_{g \sim \mathcal{N}(0, \sigma^2 \mathbf{I})} \left[\frac{\Phi(\mu + g)}{\Phi(\mu)}\right] &= - \ln \mathop{\mathbb{E}}_{a' \sim \varsigma_\mu(x)} \left[\exp\left(\frac{\sigma^2}{2} \left\|\phi(x, a')\right\|^2\right)\right] \\
&\geq - \ln \mathop{\mathbb{E}}_{a' \sim \varsigma_\mu(x)} \left[\exp\left(\frac{\sigma^2 B^2}{2}\right)\right] \\
&= - \frac{\sigma^2 B^2}{2}. \qquad (22)
\end{aligned}$$

The inequality follows from the assumption that $\|\phi(x, a')\| \leq B$. Finally, combining Equations 18 to 20 and 22, we have

$$\begin{aligned}
\pi_{\mathbb{Q}}(a \,|\, x) &\geq \exp\left(\mathop{\mathbb{E}}_{w\sim\mathcal{N}(\mu,\sigma^2\mathbf{I})}\left[w \cdot \phi(x,a) - \ln \Phi(w)\right]\right) \\
&= \exp\left(\mu \cdot \phi(x,a) - \ln\Phi(\mu) - \mathop{\mathbb{E}}_{g\sim\mathcal{N}(0,\sigma^2\mathbf{I})}\left[\ln\left(\frac{\Phi(\mu+g)}{\Phi(\mu)}\right)\right]\right) \\
&\geq \exp\left(\mu \cdot \phi(x,a) - \ln\Phi(\mu) - \ln\mathop{\mathbb{E}}_{g\sim\mathcal{N}(0,\sigma^2\mathbf{I})}\left[\frac{\Phi(\mu+g)}{\Phi(\mu)}\right]\right) \\
&\geq \exp\left(\mu \cdot \phi(x,a) - \ln\Phi(\mu) - \frac{\sigma^2 B^2}{2}\right) \\
&= \varsigma_\mu(a \,|\, x)\exp\left(-\frac{\sigma^2 B^2}{2}\right).
\end{aligned}$$

To prove the upper bound, first observe that

$$\begin{aligned}
\varsigma_w(a \,|\, x) &= \exp\left(\mu \cdot \phi(x,a) - \ln\Phi(\mu) + g \cdot \phi(x,a) - \ln\left(\frac{\Phi(\mu+g)}{\Phi(\mu)}\right)\right) \\
&= \varsigma_\mu(a \,|\, x)\exp\left(g \cdot \phi(x,a) - \ln\left(\frac{\Phi(\mu+g)}{\Phi(\mu)}\right)\right) \\
&= \varsigma_\mu(a \,|\, x)\exp\left(g \cdot \phi(x,a) - \ln\mathop{\mathbb{E}}_{a'\sim\varsigma_\mu(x)}\left[\exp(g \cdot \phi(x,a'))\right]\right) \\
&\leq \varsigma_\mu(a \,|\, x)\exp\left(g \cdot \phi(x,a) - \mathop{\mathbb{E}}_{a'\sim\varsigma_\mu(x)}\left[g \cdot \phi(x,a')\right]\right) \\
&\leq \varsigma_\mu(a \,|\, x)\mathop{\mathbb{E}}_{a'\sim\varsigma_\mu(x)}\left[\exp\left(g \cdot (\phi(x,a) - \phi(x,a'))\right)\right].
\end{aligned}$$

The inequalities follow from Jensen's inequality. We then have that

$$\begin{aligned}
\pi_{\mathbb{Q}}(a \,|\, x) &= \mathop{\mathbb{E}}_{w\sim\mathcal{N}(\mu,\sigma^2\mathbf{I})}\left[\varsigma_w(a \,|\, x)\right] \\
&\leq \varsigma_\mu(a \,|\, x)\mathop{\mathbb{E}}_{a'\sim\varsigma_\mu(x)}\mathop{\mathbb{E}}_{g\sim\mathcal{N}(0,\sigma^2\mathbf{I})}\left[\exp\left(g \cdot (\phi(x,a) - \phi(x,a'))\right)\right].
\end{aligned}$$

The right-hand inner expectation is the moment-generating function of a multivariate Gaussian:

$$\begin{aligned}
\mathop{\mathbb{E}}_{g\sim\mathcal{N}(0,\sigma^2\mathbf{I})}\left[\exp\left(g \cdot (\phi(x,a) - \phi(x,a'))\right)\right] &= \exp\left(\frac{\sigma^2}{2}\left\|\phi(x,a) - \phi(x,a')\right\|^2\right) \\
&\leq \exp\left(\frac{\sigma^2}{2}(\|\phi(x,a)\| + \|\phi(x,a')\|)^2\right) \\
&\leq \exp\left(\frac{\sigma^2}{2}(B+B)^2\right) \\
&= \exp(2\sigma^2 B^2).
\end{aligned}$$

The first inequality follows from the triangle inequality. Therefore,

$$\pi_{\mathbb{Q}}(a \,|\, x) \leq \varsigma_\mu(a \,|\, x)\mathop{\mathbb{E}}_{a'\sim\varsigma_\mu(x)}[\exp(2\sigma^2 B^2)] = \varsigma_\mu(a \,|\, x)\exp(2\sigma^2 B^2),$$

which completes the proof.

E. Proof of Theorem 2

Using Lemma 2, it is easy to show that $\hat{R}_\tau(\pi_{\mathbb{Q}}, S) \leq \hat{R}_\tau(\mu, \sigma^2, S)$. The rest of the proof follows from using Lemma 1 to upper-bound the KL divergence in Theorem 1.

F. Proofs of Propositions 1 and 2

We start by proving Proposition 1. To simplify Equation 9, we let

$$\alpha \triangleq \hat{R}_\tau(\mu, \sigma^2, S) - 1 + \frac{1}{\tau} \qquad \text{and} \qquad \beta \triangleq \frac{\Gamma(\mu_0, \sigma_0^2, \mu, \sigma^2) + 2 \ln \frac{n}{\delta}}{\tau\,(n-1)}.$$

Noting that $\hat{R}_\tau(\mu, \sigma^2, S) \leq \alpha$ (since $\tau^{-1} - 1 \geq 0$), we can upper-bound Equation 9 as

$$R(\pi_{\mathbb{Q}}) \leq \alpha + \sqrt{\alpha\beta} + \beta. \tag{23}$$

The middle term is the geometric mean of α and β, which is at most the arithmetic mean:

$$\alpha + \sqrt{\alpha\beta} + \beta \leq \alpha + \frac{\alpha + \beta}{2} + \beta = \frac{3(\alpha + \beta)}{2}. \tag{24}$$

We therefore obtain an upper bound on Equation 9 that omits the middle term, which can be tricky to optimize due to the interaction between α and β. If we optimize this upper bound,

$$\begin{aligned}
\operatorname*{arg\,min}_{\substack{\mu \in \mathbb{R}^d \\ \sigma^2 \in (0, \sigma_0^2]}} \frac{3(\alpha + \beta)}{2} &= \operatorname*{arg\,min}_{\substack{\mu \in \mathbb{R}^d \\ \sigma^2 \in (0, \sigma_0^2]}} \alpha + \beta \\
&= \operatorname*{arg\,min}_{\substack{\mu \in \mathbb{R}^d \\ \sigma^2 \in (0, \sigma_0^2]}} \hat{R}_\tau(\mu, \sigma^2, S) - 1 + \frac{1}{\tau} + \frac{\Gamma(\mu_0, \sigma_0^2, \mu, \sigma^2) + 2 \ln \frac{n}{\delta}}{\tau\,(n-1)} \\
&= \operatorname*{arg\,min}_{\substack{\mu \in \mathbb{R}^d \\ \sigma^2 \in (0, \sigma_0^2]}} \hat{R}_\tau(\mu, \sigma^2, S) + \frac{\Gamma(\mu_0, \sigma_0^2, \mu, \sigma^2)}{\tau\,(n-1)} \\
&= \operatorname*{arg\,min}_{\substack{\mu \in \mathbb{R}^d \\ \sigma^2 \in (0, \sigma_0^2]}} \hat{R}_\tau(\mu, \sigma^2, S) + \frac{\frac{1}{\sigma_0^2} \left\| \mu - \mu_0 \right\|^2 + d \ln \frac{\sigma_0^2}{\sigma^2}}{\tau\,(n-1)} \\
&= \operatorname*{arg\,min}_{\substack{\mu \in \mathbb{R}^d \\ \sigma^2 \in (0, \sigma_0^2]}} \hat{R}_\tau(\mu, \sigma^2, S) + \frac{\frac{1}{\sigma_0^2} \left\| \mu - \mu_0 \right\|^2 - d \ln \sigma^2}{\tau\,(n-1)},
\end{aligned}$$

we obtain Equation 11.

To prove Proposition 2, we upper-bound $\hat{R}_\tau(\mu, \sigma^2, S)$ by using the fact that $u \ln v \leq uv$ for $u, v \geq 0$. Setting

$$u_i \triangleq \frac{r_i}{\max\{p_i, \tau\}} \qquad \text{and} \qquad v_i \triangleq \frac{\varsigma_\mu(a_i \,|\, x_i)}{\exp(\frac{\sigma^2 B^2}{2})},$$

we have that

$$\begin{aligned}
\hat{R}_\tau(\mu, \sigma^2, S) - 1 &= -\frac{1}{n} \sum_{i=1}^{n} \frac{r_i}{\max\{p_i, \tau\}} \frac{\varsigma_\mu(a_i \,|\, x_i)}{\exp(\frac{\sigma^2 B^2}{2})} \\
&= -\frac{1}{n} \sum_{i=1}^{n} u_i v_i \\
&\leq -\frac{1}{n} \sum_{i=1}^{n} u_i \ln v_i.
\end{aligned}$$

Let

$$\gamma \triangleq \frac{1}{\tau} - \frac{1}{n} \sum_{i=1}^{n} u_i \ln v_i,$$

and observe that $\alpha \leq \gamma$. Thus, by Equations 23 and 24,

$$R(\pi_{\mathbb{Q}}) \leq \frac{3(\alpha+\beta)}{2} \leq \frac{3(\gamma+\beta)}{2}.$$

Optimizing this upper bound yields the following equivalence:

$$\begin{aligned}
\operatorname*{arg\,min}_{\substack{\mu\in\mathbb{R}^d \\ \sigma^2\in(0,\sigma_0^2]}} \frac{3(\gamma+\beta)}{2} &= \operatorname*{arg\,min}_{\substack{\mu\in\mathbb{R}^d \\ \sigma^2\in(0,\sigma_0^2]}} \gamma+\beta \\
&= \operatorname*{arg\,min}_{\substack{\mu\in\mathbb{R}^d \\ \sigma^2\in(0,\sigma_0^2]}} \frac{1}{\tau} + \frac{1}{n}\sum_{i=1}^{n} -u_i \ln v_i + \frac{\Gamma(\mu_0,\sigma_0^2,\mu,\sigma^2) + 2\ln\frac{n}{\delta}}{\tau\,(n-1)} \\
&= \operatorname*{arg\,min}_{\substack{\mu\in\mathbb{R}^d \\ \sigma^2\in(0,\sigma_0^2]}} \frac{1}{n}\sum_{i=1}^{n} -u_i \ln v_i + \frac{\|\mu-\mu_0\|^2}{\sigma_0^2\,\tau\,(n-1)} - \frac{d\ln\sigma^2}{\tau\,(n-1)} \\
&= \operatorname*{arg\,min}_{\substack{\mu\in\mathbb{R}^d \\ \sigma^2\in(0,\sigma_0^2]}} \frac{1}{n}\sum_{i=1}^{n} -\frac{r_i \ln \varsigma_\mu(a_i \,|\, x_i)}{\max\{p_i,\tau\}} + \frac{r_i\sigma^2 B^2}{2\max\{p_i,\tau\}} + \frac{\|\mu-\mu_0\|^2}{\sigma_0^2\tau(n-1)} - \frac{d\ln\sigma^2}{\tau(n-1)}.
\end{aligned}$$

Observe that μ and σ^2 never interact multiplicatively in the objective function. We can therefore solve each sub-optimization separately.

Starting with μ, we simply isolate the relevant terms and obtain Equation 12. For σ^2, we must solve

$$\operatorname*{arg\,min}_{\sigma^2\in(0,\sigma_0^2]} \frac{1}{n}\sum_{i=1}^{n} \frac{r_i B^2\sigma^2}{2\max\{p_i,\tau\}} - \frac{d\ln\sigma^2}{\tau\,(n-1)}.$$

Note that this objective is convex in σ^2. If we ignore the constraint that $\sigma^2 \in (0,\sigma_0^2]$ and let σ^2 be any real number, then the problem has an analytic solution:

$$\operatorname*{arg\,min}_{\sigma^2\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^{n} \frac{r_i B^2\sigma^2}{2\max\{p_i,\tau\}} - \frac{d\ln\sigma^2}{\tau(n-1)} = \frac{2d}{B^2\tau(n-1)}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{\max\{p_i,\tau\}}\right)^{-1}.$$

This can be verified by setting the derivative equal to 0 and solving for σ^2. Suppose the solution to the unconstrained problem lies outside of the feasible region for the constrained problem, $(0,\sigma_0^2]$. It is easily verified that the unconstrained solution is strictly positive; thus, it must be greater than σ_0^2. Since the objective function is convex, we must then have that the solution to the constrained problem lies at the upper boundary, σ_0^2, which is the closest point to the unconstrained solution. Thus, the minimizer of the constrained problem is either the unconstrained solution or σ_0^2; whichever one is smaller.

G. Connection to Policy Gradient Methods

Those familiar with reinforcement learning may see connections between Equation 12 and *policy gradient* methods. By the policy gradient theorem (Sutton et al., 2000), the gradient of the expected reward[3] is precisely the expected, reward-weighted gradient of the log-likelihood,

$$\nabla \mathop{\mathbb{E}}_{a\sim\pi(x)}[\rho(x,a)] = \mathop{\mathbb{E}}_{a\sim\pi(x)}[\rho(x,a)\nabla\ln\pi(a\,|\,x)].$$

In online, on-policy training, the expectation is typically approximated by sampling actions from the policy. In offline, off-policy training, the expectation can be approximated by samples from the logging policy, with importance weight $\pi(a\,|\,x)/\pi_0(a\,|\,x)$ to counteract bias. We then obtain a gradient that looks like the gradient of Equation 12, albeit weighted by $\pi(a\,|\,x)$ and without the regularization term.

[3] In reinforcement learning, the expectation would be over trajectories, which we omit for simplicity.

H. Proof of Theorem 3

To prove Theorem 3, we start by borrowing a result from Liu et al. (2017), which we simplify and specialize for our use case.

Lemma 4 ((Liu et al., 2017, Lemma 1)). *Let $D_{\text{H}}(S, S')$ denote the Hamming distance between two datasets, S, S'. Suppose there exists a constant, $\alpha > 0$, such that*

$$\sup_{S,S':D_{\text{H}}(S,S')=1} \|\hat{\mu}_0(S) - \hat{\mu}_0(S')\| \leq \alpha. \tag{25}$$

(In other words, perturbing any single training example can change the learned parameters by at most α.) Then, for any $\delta \in (0, 1)$,

$$\Pr_{S\sim(\mathbb{D}\times\pi_0)^n} \left\{ \|\hat{\mu}_0(S) - \bar{\mu}_0\| \geq \alpha\sqrt{2n \ln \frac{2}{\delta}} \right\} \leq \delta.$$

To apply Lemma 4, we must identify a value of α that satisfies Equation 25.

Lemma 5. *If the loss function, L, is convex and β-Lipschitz with respect to its first argument, then the minimizer, $\hat{\mu}_0(S)$, satisfies Equation 25 for $\alpha = \frac{\beta}{\lambda n}$.*

Proof. Without loss of generality, assume that the index of the example at which S and S' differ is i. It easily verified that the regularizer, $\lambda \|w\|^2$, is (2λ)-strongly convex; and since L is assumed to be convex, the regularized objective, F (Equation 13), is also (2λ)-strongly convex. Therefore, using the definition of strongly convex functions, and the symmetry of distances, we have that

$$\begin{aligned}
\|\hat{\mu}_0(S) - \hat{\mu}_0(S')\|^2 &= \frac{1}{2}\|\hat{\mu}_0(S) - \hat{\mu}_0(S')\|^2 + \frac{1}{2}\|\hat{\mu}_0(S') - \hat{\mu}_0(S)\|^2 \\
&\leq \frac{1}{2\lambda}\left(F(\hat{\mu}_0(S), S') - F(\hat{\mu}_0(S'), S')\right) + \frac{1}{2\lambda}\left(F(\hat{\mu}_0(S'), S) - F(\hat{\mu}_0(S), S)\right) \\
&= \frac{1}{2\lambda}\left(F(\hat{\mu}_0(S'), S) - F(\hat{\mu}_0(S'), S')\right) + \frac{1}{2\lambda}\left(F(\hat{\mu}_0(S), S') - F(\hat{\mu}_0(S), S)\right) \\
&= \frac{1}{2\lambda n}\left(L(\hat{\mu}_0(S'), x_i, a_i) - L(\hat{\mu}_0(S'), x'_i, a'_i)\right) + \frac{1}{2\lambda n}\left(L(\hat{\mu}_0(S), x'_i, a'_i) - L(\hat{\mu}_0(S), x_i, a_i)\right) \\
&= \frac{1}{2\lambda n}\left(L(\hat{\mu}_0(S'), x_i, a_i) - L(\hat{\mu}_0(S), x_i, a_i)\right) + \frac{1}{2\lambda n}\left(L(\hat{\mu}_0(S), x'_i, a'_i) - L(\hat{\mu}_0(S'), x'_i, a'_i)\right) \\
&\leq \frac{\beta}{2\lambda n}\left(\|\hat{\mu}_0(S') - \hat{\mu}_0(S)\| + \|\hat{\mu}_0(S) - \hat{\mu}_0(S')\|\right) \\
&= \frac{\beta}{\lambda n}\|\hat{\mu}_0(S) - \hat{\mu}_0(S')\|.
\end{aligned}$$

Dividing each side by $\|\hat{\mu}_0(S) - \hat{\mu}_0(S')\|$ completes the proof. □

Now, we can apply Lemma 4 to show that $\hat{\mu}_0(S)$ concentrates around $\bar{\mu}_0$.

Lemma 6. *If the loss function, L, is convex and β-Lipschitz with respect to its first argument, then for any $\delta \in (0, 1)$,*

$$\Pr_{S\sim(\mathbb{D}\times\pi_0)^n} \left\{ \|\hat{\mu}_0(S) - \bar{\mu}_0\| \geq \frac{\beta}{\lambda}\sqrt{\frac{2\ln\frac{2}{\delta}}{n}} \right\} \leq \delta.$$

Proof. Follows immediately from Lemmas 4 and 5, with $\alpha = \frac{\beta}{\lambda n}$. □

We are now ready to prove Theorem 3. We start by applying Theorem 2, with μ_0 replaced by $\bar{\mu}_0$, and δ replaced by $\delta/2$. With probability at least $1 - \delta/2$,

$$R(\pi_{\mathbb{Q}}) \leq \hat{R}_\tau(\mu, \sigma^2, S) + \sqrt{\frac{\left(\hat{R}_\tau(\mu, \sigma^2, S) - 1 + \frac{1}{\tau}\right)\left(\Gamma(\bar{\mu}_0, \sigma_0^2, \mu, \sigma^2) + 2\ln\frac{2n}{\delta}\right)}{\tau\,(n-1)}} + \frac{\left(\Gamma(\bar{\mu}_0, \sigma_0^2, \mu, \sigma^2) + 2\ln\frac{2n}{\delta}\right)}{\tau\,(n-1)}.$$

Then, using the triangle inequality and Lemma 6, we have that

$$\begin{aligned}\|\mu - \bar{\mu}_0\| &\leq \|\mu - \hat{\mu}_0(S)\| + \|\hat{\mu}_0(S) - \bar{\mu}_0\| \\ &\leq \|\mu - \hat{\mu}_0(S)\| + \frac{\beta}{\lambda}\sqrt{\frac{2\ln\frac{4}{\delta}}{n}},\end{aligned}$$

with probability at least $1 - \delta/2$. Substituting this into Equation 10 yields

$$\begin{aligned}\Gamma(\bar{\mu}_0, \sigma_0^2, \mu, \sigma^2) &= \frac{\|\mu - \bar{\mu}_0\|^2}{\sigma_0^2} + d\ln\frac{\sigma_0^2}{\sigma^2} \\ &\leq \frac{\left(\|\mu - \hat{\mu}_0(S)\| + \frac{\beta}{\lambda}\sqrt{\frac{2\ln\frac{4}{\delta}}{n}}\right)^2}{\sigma_0^2} + d\ln\frac{\sigma_0^2}{\sigma^2} \\ &= \hat{\Gamma}(\hat{\mu}_0(S), \sigma_0^2, \mu, \sigma^2),\end{aligned}$$

with probability at least $1 - \delta/2$. Thus, Equation 15 holds with probability at least $1 - \delta$.

References

T. Liu, G. Lugosi, G. Neu, and D. Tao. Algorithmic stability and hypothesis complexity. In *International Conference on Machine Learning*, 2017.

D. McAllester. Simplified PAC-Bayesian margin bounds. In *COLT*, pages 203–215, 2003.

R. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Neural Information Processing Systems*, 2000.

A. Swaminathan and T. Joachims. The self-normalized estimator for counterfactual learning. In *Neural Information Processing Systems*, 2015.