Supplementary material

A. Implementation Details for SparseMAP Solvers

A.1. Conditional Gradient Variants

We adapt the presentation of vanilla, away-step and pairwise conditional gradient of Lacoste-Julien & Jaggi (2015).

Recall the SparseMAP optimization problem (Equation 5), which we rewrite below as a minimization, to align with the formulation in (Lacoste-Julien & Jaggi, 2015)

$$\mathsf{SparseMAP}_{\boldsymbol{A}}(\boldsymbol{\eta}) \coloneqq \operatorname*{arg\,min}_{\boldsymbol{u},\ [\boldsymbol{u},\boldsymbol{v}]\in\mathcal{M}_{\boldsymbol{A}}} f(\boldsymbol{u},\boldsymbol{v}), \qquad \text{where } f(\boldsymbol{u},\boldsymbol{v}) \coloneqq \frac{1}{2}\left\|\boldsymbol{u}\right\|_2^2 - \boldsymbol{\eta}_U^\top \boldsymbol{u} - \boldsymbol{\eta}_F^\top \boldsymbol{v}.$$

The gradients of the objective function f w.r.t. the two variables are

$$\nabla_{\boldsymbol{u}} f(\boldsymbol{u}', \boldsymbol{v}') = \boldsymbol{u}' - \boldsymbol{\eta}_U, \qquad \nabla_{\boldsymbol{v}} f(\boldsymbol{u}', \boldsymbol{v}') = -\boldsymbol{\eta}_V.$$

The ingredients required to apply conditional gradient algorithms are solving linear minimization problem, selecting the away step, computing the Wolfe gap, and performing line search.

Linear minimization problem. For SparseMAP, this amounts to a MAP inference call, since

$$\begin{aligned}
&\operatorname*{arg\,min}_{[\boldsymbol{u},\boldsymbol{v}]\in\mathcal{M}_{\boldsymbol{A}}} \left\langle \nabla_{\boldsymbol{u}} f(\boldsymbol{u'},\boldsymbol{v'}), \boldsymbol{u}\right\rangle + \left\langle \nabla_{\boldsymbol{v}} f(\boldsymbol{u'},\boldsymbol{v'}), \boldsymbol{v}\right\rangle \\
=&\operatorname*{arg\,min}_{[\boldsymbol{u},\boldsymbol{v}]\in\mathcal{M}_{\boldsymbol{A}}} (\boldsymbol{u}' - \boldsymbol{\eta}_U)^\top \boldsymbol{u} - \boldsymbol{\eta}_F^\top \boldsymbol{v} \\
=&\{[\boldsymbol{m}_s; \boldsymbol{n}_s] \ :\ s \in \mathsf{MAP}_{\boldsymbol{A}}(\boldsymbol{\eta}_U - \boldsymbol{u}', \boldsymbol{\eta}_F)\}.
\end{aligned}$$

where we assume $\mathsf{MAP}_{\boldsymbol{A}}$ yields the set of maximally-scoring structures.

Away step selection. This step involves searching the currently selected structures in the active set $\mathcal{I}$ with the *opposite* goal: finding the structure *maximizing* the linearization

$$\begin{aligned}
&\operatorname*{arg\,max}_{s\in\mathcal{I}} \left\langle \nabla_{\boldsymbol{u}} f(\boldsymbol{u'},\boldsymbol{v'}), \boldsymbol{m}_s\right\rangle + \left\langle \nabla_{\boldsymbol{v}} f(\boldsymbol{u'},\boldsymbol{v'}), \boldsymbol{n}_s\right\rangle \\
=&\operatorname*{arg\,max}_{s\in\mathcal{I}} (\boldsymbol{u}' - \boldsymbol{\eta}_U)^\top \boldsymbol{m}_s - \boldsymbol{\eta}_F^\top \boldsymbol{n}_s
\end{aligned}$$

Wolfe gap. The gap at a point $\boldsymbol{d} = [\boldsymbol{d}_{\boldsymbol{u}}; \boldsymbol{d}_{\boldsymbol{v}}]$ is given by

$$\begin{aligned}
\operatorname{gap}(\boldsymbol{d}, \boldsymbol{u}') &\coloneqq \left\langle -\nabla_{\boldsymbol{u}} f(\boldsymbol{u'},\boldsymbol{v'}), \boldsymbol{d}_{\boldsymbol{u}}\right\rangle + \left\langle -\nabla_{\boldsymbol{v}} f(\boldsymbol{u'},\boldsymbol{v'}), \boldsymbol{d}_{\boldsymbol{v}}\right\rangle \\
&= \left\langle \boldsymbol{\eta}_U - \boldsymbol{u}', \boldsymbol{d}_{\boldsymbol{u}}\right\rangle + \left\langle \boldsymbol{\eta}_F, \boldsymbol{d}_{\boldsymbol{v}}\right\rangle.
\end{aligned} \tag{7}$$

Line search. Once we have picked a direction $\boldsymbol{d} = [\boldsymbol{d}_{\boldsymbol{u}}; \boldsymbol{d}_{\boldsymbol{v}}]$, we can pick the optimal step size by solving a simple optimization problem. Let $\boldsymbol{u}_\gamma \coloneqq \boldsymbol{u}' + \gamma \boldsymbol{d}_{\boldsymbol{u}}$, and $\boldsymbol{v}_\gamma \coloneqq \boldsymbol{v}' + \gamma \boldsymbol{d}_{\boldsymbol{v}}$. We seek γ so as to optimize

$$\operatorname*{arg\,min}_{\gamma\in[0,\gamma_{\max}]} f(\boldsymbol{u}_\gamma, \boldsymbol{v}_\gamma)$$

Setting the gradient w.r.t. γ to 0 yields

$$\begin{aligned}
0 &= \frac{\partial}{\partial\gamma} f(\boldsymbol{u}_\gamma, \boldsymbol{v}_\gamma) \\
&= \left\langle \boldsymbol{d}_{\boldsymbol{u}}, \nabla_{\boldsymbol{u}} f(\boldsymbol{u}_\gamma, \boldsymbol{v}_\gamma)\right\rangle + \left\langle \boldsymbol{d}_{\boldsymbol{v}}, \nabla_{\boldsymbol{v}} f(\boldsymbol{u}_\gamma, \boldsymbol{v}_\gamma)\right\rangle \\
&= \left\langle \boldsymbol{d}_{\boldsymbol{u}}, \boldsymbol{u}' + \gamma\boldsymbol{d}_{\boldsymbol{u}} - \boldsymbol{\eta}_U\right\rangle + \left\langle \boldsymbol{d}_{\boldsymbol{v}}, -\boldsymbol{\eta}_F\right\rangle \\
&= \gamma \left\|\boldsymbol{d}_{\boldsymbol{u}}\right\|_2^2 + \boldsymbol{u}'^\top \boldsymbol{d}_{\boldsymbol{u}} - \boldsymbol{\eta}^\top \boldsymbol{d}
\end{aligned}$$

We may therefore compute the optimal step size γ as

$$\gamma = \mathsf{max}\left(0, \mathsf{min}\left(\gamma_{\mathsf{max}}, \frac{\boldsymbol{\eta}^\top \boldsymbol{d} - \boldsymbol{u}'^\top \boldsymbol{d}_{\boldsymbol{u}}}{\|\boldsymbol{d}_{\boldsymbol{u}}\|_2^2}\right)\right) \tag{8}$$

Algorithm 1 Conditional gradient for SparseMAP

Initialization: $s^{(0)} \leftarrow \mathsf{MAP}_{\boldsymbol{A}}(\boldsymbol{\eta}_U, \boldsymbol{\eta}_F)$; $\mathcal{I}^{(0)} = \{s^{(0)}\}$; $\boldsymbol{y}^{(0)} = \boldsymbol{e}_{s^{(0)}}$; $[\boldsymbol{u}^{(0)}; \boldsymbol{v}^{(0)}] = \boldsymbol{a}_{s^{(0)}}$
for $t = 0 \dots t_{\mathsf{max}}$ **do**
  $s \leftarrow \mathsf{MAP}_{\boldsymbol{A}}(\boldsymbol{\eta}_U - \boldsymbol{u}^{(t)}, \boldsymbol{\eta}_F)$; $\boldsymbol{d}^{\mathrm{F}} \leftarrow \boldsymbol{a}_s - [\boldsymbol{u}^{(t)}; \boldsymbol{v}^{(t)}]$ (forward direction)
  $w \leftarrow \underset{w \in \mathcal{I}^{(t)}}{\mathsf{arg\,max}} \ (\boldsymbol{\eta}_U - \boldsymbol{u}^{(t)})^\top \boldsymbol{m}_w + \boldsymbol{\eta}_F^\top \boldsymbol{n}_w$; $\boldsymbol{d}^{\mathrm{W}} \leftarrow [\boldsymbol{u}^{(t)}; \boldsymbol{v}^{(t)}] - \boldsymbol{a}_w$ (away direction)
  if $\mathrm{gap}(\boldsymbol{d}^{\mathrm{F}}, \boldsymbol{u}^{(t)}) < \epsilon$ **then**
    return $\boldsymbol{u}^{(t)}$ (Equation 7)
  end if
  if variant $=$ vanilla **then**
    $\boldsymbol{d} \leftarrow \boldsymbol{d}^{\mathrm{F}}$; $\gamma_{\mathsf{max}} \leftarrow 1$
  else if variant $=$ pairwise **then**
    $\boldsymbol{d} \leftarrow \boldsymbol{d}^{\mathrm{F}} + \boldsymbol{d}^{\mathrm{W}}$; $\gamma_{\mathsf{max}} \leftarrow y_w$
  else if variant $=$ away-step **then**
    if $\mathrm{gap}(\boldsymbol{d}^{\mathrm{F}}, \boldsymbol{u}^{(t)}) \geq \mathrm{gap}(\boldsymbol{d}^{\mathrm{W}}, \boldsymbol{u}^{(t)})$ **then**
      $\boldsymbol{d} \leftarrow \boldsymbol{d}^{\mathrm{F}}$; $\gamma_{\mathsf{max}} \leftarrow 1$
    else
      $\boldsymbol{d} \leftarrow \boldsymbol{d}^{\mathrm{A}}$; $\gamma_{\mathsf{max}} \leftarrow y_w/(1 - y_w)$
    end if
  end if
  Compute step size γ (Equation 8)
  $[\boldsymbol{u}^{(t+1)}; \boldsymbol{v}^{(t+1)}] \leftarrow [\boldsymbol{u}^{(t)}; \boldsymbol{v}^{(t)}] + \boldsymbol{d}$
  Update $\mathcal{I}^{(t+1)}$ and $\boldsymbol{y}^{(t+1)}$ accordingly.
end for

A.2. The Active Set Algorithm

We use a variant of the active set algorithm (Nocedal & Wright, 1999, Ch. 16.4 & 16.5) as proposed for the quadratic subproblems of the AD^3 algorithm; our presentation follows (Martins et al., 2015, Algorithm 3). At each step, the active set algorithm solves a relaxed variant of the SparseMAP QP, relaxing the non-negativity constraint on $\boldsymbol{y}$, and restricting the solution to the current active set $\mathcal{I}$

$$\text{minimize}_{\boldsymbol{y}_\mathcal{I} \in \mathbb{R}^{|\mathcal{I}|}} \quad \frac{1}{2}\|\boldsymbol{M}_\mathcal{I}\boldsymbol{y}_\mathcal{I}\|_2^2 - \boldsymbol{\eta}^\top \boldsymbol{A}_\mathcal{I}\boldsymbol{y}_\mathcal{I} \qquad \text{subject to} \quad \mathbf{1}^\top \boldsymbol{y}_\mathcal{I} = 1$$

whose solution can be found by solving the KKT system

$$\begin{bmatrix} \boldsymbol{M}_\mathcal{I}^\top \boldsymbol{M}_\mathcal{I} & \mathbf{1} \\ \mathbf{1}^\top & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{y}_\mathcal{I} \\ \tau \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}_\mathcal{I}^\top \boldsymbol{\eta} \\ 1 \end{bmatrix}. \tag{9}$$

At each iteration, the (symmetric) design matrix in Equation 9 is updated by adding or removing a row and a column; therefore its inverse (or a decomposition) may be efficiently maintained and updated.

Line search. The optimal step size for moving a feasible current estimate $\boldsymbol{y}'$ toward a solution $\hat{\boldsymbol{y}}$ of Equation 9, while keeping feasibility, is given by (Martins et al., 2015, Equation 31)

$$\gamma = \mathsf{min}\left(1, \min_{s \in \mathcal{I},\ y'_s > \hat{y}_s} \frac{y'_s}{y'_s - \hat{y}_s}\right) \tag{10}$$

When $\gamma \leq 1$ this update zeros out a coordinate of $\boldsymbol{y}'$; otherwise, $\mathcal{I}$ remains the same.

Algorithm 2 Active Set algorithm for SparseMAP

Initialization: $s^{(0)} \leftarrow \mathsf{MAP}_{\boldsymbol{A}}(\boldsymbol{\eta}_U, \boldsymbol{\eta}_F)$; $\mathcal{I}^{(0)} = \{s^{(0)}\}$; $\boldsymbol{y}^{(0)} = \boldsymbol{e}_{s^{(0)}}$; $[\boldsymbol{u}^{(0)}; \boldsymbol{v}^{(0)}] = \boldsymbol{a}_{s^{(0)}}$
for $t = 0 \dots t_{\mathsf{max}}$ **do**
  Solve the relaxed QP restricted to $\mathcal{I}^{(t)}$; get $\hat{\boldsymbol{y}}, \hat{\tau}, \hat{\boldsymbol{u}} = \boldsymbol{M}\hat{\boldsymbol{y}}$ (Equation 9)
  if $\hat{\boldsymbol{y}} = \boldsymbol{y}^{(t)}$ **then**
    $s \leftarrow \mathsf{MAP}_{\boldsymbol{A}}(\boldsymbol{\eta}_U - \hat{\boldsymbol{u}}, \boldsymbol{\eta}_F)$
    if $\mathrm{gap}(\boldsymbol{a}_s, \hat{\boldsymbol{u}}) \leq \hat{\tau}$ **then**
      return $\boldsymbol{u}^{(t)}$ (Equation 7)
    else
      $\mathcal{I}^{(t+1)} \leftarrow \mathcal{I}^{(t)} \cup \{s\}$
    end if
  else
    Compute step size γ (Equation 10)
    $\boldsymbol{y}^{(t+1)} \leftarrow (1-\gamma)\boldsymbol{y}^{(t)} + \gamma\hat{\boldsymbol{y}}$ (sparse update)
    Update $\mathcal{S}^{(t+1)}$ if necessary
  end if
end for

B. Computing the SparseMAP Jacobian: Proof of Proposition 1

Recall that SparseMAP is defined as the $\boldsymbol{u}^\star$ that maximizes the value of the quadratic program (Equation 5),

$$g(\boldsymbol{\eta}_U, \boldsymbol{\eta}_F) \coloneqq \max_{[\boldsymbol{u};\boldsymbol{v}] \in \mathcal{M}_{\boldsymbol{A}}} \boldsymbol{\eta}_U^\top \boldsymbol{u} + \boldsymbol{\eta}_F^\top \boldsymbol{v} - \frac{1}{2}\|\boldsymbol{u}\|_2^2 . \tag{11}$$

As the ℓ_2^2 norm is strongly convex, there is always a unique minimizer $\boldsymbol{u}^\star$ (implying that SparseMAP is well-defined), and the convex conjugate of the QP in (11), $g^*(\boldsymbol{u}, \boldsymbol{v}) = \left\{\frac{1}{2}\|\boldsymbol{u}\|_2^2, [\boldsymbol{u};\boldsymbol{v}] \in \mathcal{M}_{\boldsymbol{A}}; -\infty \text{ otherwise}\right\}$ is smooth in $\boldsymbol{u}$, implying that SparseMAP (which only returns $\boldsymbol{u}$) is Lipschitz-continuous and thus differentiable almost everywhere.

We now rewrite the QP in Equation 11 in terms of the convex combination of vertices of the marginal polytope

$$\min_{\boldsymbol{y} \in \triangle^D} \frac{1}{2}\|\boldsymbol{M}\boldsymbol{y}\|_2^2 - \boldsymbol{\theta}^\top \boldsymbol{y} \qquad \text{where } \boldsymbol{\theta} \coloneqq \boldsymbol{A}^\top \boldsymbol{\eta} \tag{12}$$

We use the optimality conditions of problem 12 to derive an explicit relationship between $\boldsymbol{u}^\star$ and $\boldsymbol{x}$. At an optimum, the following KKT conditions hold

$$\boldsymbol{M}^\top \boldsymbol{M}\boldsymbol{y}^\star - \boldsymbol{\lambda}^\star + \tau^\star \mathbf{1} = \boldsymbol{\theta} \tag{13}$$
$$\mathbf{1}^\top \boldsymbol{y}^\star = 1 \tag{14}$$
$$\boldsymbol{y}^\star \geq \mathbf{0} \tag{15}$$
$$\boldsymbol{\lambda}^\star \geq \mathbf{0} \tag{16}$$
$$\boldsymbol{\lambda}^{\star\top} \boldsymbol{y}^\star = 0 \tag{17}$$

Let $\mathcal{I}$ denote the support of $\boldsymbol{y}^\star$, *i.e.*, $\mathcal{I} = \{s \,:\, y_s^\star > 0\}$. From Equation 17 we have $\boldsymbol{\lambda}_{\mathcal{I}} = \mathbf{0}$ and therefore

$$\boldsymbol{M}_{\mathcal{I}}{}^\top \boldsymbol{M}_{\mathcal{I}} \boldsymbol{y}_{\mathcal{I}}^\star + \tau^\star \mathbf{1} = \boldsymbol{\theta}_{\mathcal{I}} \tag{18}$$
$$\mathbf{1}^\top \boldsymbol{y}_{\mathcal{I}}^\star = 1 \tag{19}$$

Solving for $\boldsymbol{y}_{\mathcal{I}}^\star$ in Equation 18 we get a direct expression

$$\boldsymbol{y}_{\mathcal{I}}{}^\star = (\boldsymbol{M}_{\mathcal{I}}{}^\top \boldsymbol{M}_{\mathcal{I}})^{-1}(\boldsymbol{\theta}_{\mathcal{I}} - \tau^\star \mathbf{1}) = \boldsymbol{Z}(\boldsymbol{\theta}_{\mathcal{I}} - \tau^\star \mathbf{1}).$$

where we introduced $\boldsymbol{Z} = (\boldsymbol{M}^\top \boldsymbol{M})^{-1}$. Solving for $\tau^\star$ yields

$$\tau^\star = \frac{1}{\mathbf{1}^T \boldsymbol{Z} \mathbf{1}} \left(\mathbf{1}^T \boldsymbol{Z} \boldsymbol{\theta}_\mathcal{I} - 1\right)$$

Plugging this back and left-multiplying by $\boldsymbol{M}_\mathcal{I}$ we get

$$\boldsymbol{u}^\star = \boldsymbol{M}_\mathcal{I} \boldsymbol{y}^\star_\mathcal{I} = \boldsymbol{M}_\mathcal{I} \boldsymbol{Z} \left(\boldsymbol{\theta}_\mathcal{I} - \frac{1}{\mathbf{1}^\top \boldsymbol{Z} \mathbf{1}} \mathbf{1}^\top \boldsymbol{Z} \boldsymbol{\theta}_\mathcal{I} \mathbf{1} + \frac{1}{\mathbf{1}^\top \boldsymbol{Z} \mathbf{1}} \mathbf{1}\right)$$

Note that, in a neighborhood of $\boldsymbol{\eta}$, the support of the solution $\mathcal{I}$ is constant. (On the measure-zero set of points where the support changes, SparseMAP is subdifferentiable and our assumption yields a generalized Jacobian (Clarke, 1990).) Differentiating w.r.t. the score of a configuration θ_s, we get the expression

$$\frac{\partial \boldsymbol{u}^\star}{\partial \theta_s} = \begin{cases} \boldsymbol{M} \left(\boldsymbol{I} - \frac{1}{\mathbf{1}^T \boldsymbol{Z} \mathbf{1}} \boldsymbol{Z} \mathbf{1} \mathbf{1}^T\right) \boldsymbol{z}_s & s \in \mathcal{I} \\ \mathbf{0} & s \notin \mathcal{I} \end{cases} \tag{20}$$

Since $\theta_s = \boldsymbol{a}_s^\top \boldsymbol{\eta}$, by the chain rule, we get the desired result

$$\frac{\partial \boldsymbol{u}^\star}{\partial \boldsymbol{\eta}} = \frac{\partial \boldsymbol{u}^\star}{\partial \boldsymbol{\theta}} \boldsymbol{A}^\top. \tag{21}$$

C. Fenchel-Young Losses: Proof of Proposition 2

We recall that the structured Fenchel-Young loss defined by a convex $\Omega : \mathbb{R}^D \to \mathbb{R}$ and a matrix $\boldsymbol{A}$ is defined as

$$\ell_{\Omega,\boldsymbol{A}} : \mathbb{R}^k \times \triangle^D \to \mathbb{R}, \quad \ell_{\Omega,\boldsymbol{A}}(\boldsymbol{\eta}, \boldsymbol{y}) \coloneqq \Omega^*_\triangle(\boldsymbol{A}^\top \boldsymbol{\eta}) + \Omega_\triangle(\boldsymbol{y}) - \boldsymbol{\eta}^\top \boldsymbol{A} \boldsymbol{y}.$$

Since $\Omega_\triangle$ is the restriction of a convex function to a convex set, it is convex (Boyd & Vandenberghe, 2004, Section 3.1.2).

Property 1. From the Fenchel-Young inequality (Fenchel, 1949; Boyd & Vandenberghe, 2004, Section 3.3.2), we have

$$\boldsymbol{\theta}^\top \boldsymbol{y} \leq \Omega^*_\triangle(\boldsymbol{\theta}) + \Omega_\triangle(\boldsymbol{y}).$$

In particular, when $\boldsymbol{\theta} = \boldsymbol{A}^\top \boldsymbol{\eta}$,

$$\begin{aligned} 0 &\leq -\boldsymbol{\eta}^\top \boldsymbol{A} \boldsymbol{y} + \Omega^*_\triangle(\boldsymbol{A}^\top \boldsymbol{\eta}) + \Omega_\triangle(\boldsymbol{y}) \\ &= \ell_{\Omega,\boldsymbol{A}}(\boldsymbol{\eta}, \boldsymbol{y}). \end{aligned}$$

Equality is achieved when

$$\begin{aligned} \Omega^*_\triangle(\boldsymbol{A}^\top \boldsymbol{\eta}) = \boldsymbol{\eta}^\top \boldsymbol{A} \boldsymbol{y} - \Omega_\triangle(\boldsymbol{y}) \iff \\ \max_{\boldsymbol{y}' \in \triangle^d} \boldsymbol{\eta}^\top \boldsymbol{A} \boldsymbol{y}' - \Omega(\boldsymbol{y}') = \boldsymbol{\eta}^\top \boldsymbol{A} \boldsymbol{y} - \Omega(\boldsymbol{y}), \end{aligned}$$

where we used the fact that $\boldsymbol{y} \in \triangle^d$. The second part of the claim follows.

Property 2. To prove convexity in $\boldsymbol{\eta}$, we rewrite the loss, for fixed $\boldsymbol{y}$, as

$$\ell_{\Omega,\boldsymbol{A}}(\boldsymbol{\eta}) = h(\boldsymbol{A}^\top \boldsymbol{\eta}) + \text{const}, \quad \text{where} \quad h(\boldsymbol{\theta}) = \Omega^*_\triangle(\boldsymbol{\theta}) - \boldsymbol{\theta}^\top \boldsymbol{y}.$$

$\Omega^*_\triangle$ is a convex conjugate, and thus itself convex. Linear functions are convex, and the sum of two convex functions is convex, therefore h is convex. Finally, the composition of a convex function with a linear function is convex as well, thus the function $\left(h \boldsymbol{A}^\top\right)$ is convex. Convexity of $\ell_{\Omega,\boldsymbol{A}}$ in $\boldsymbol{\eta}$ directly follows. Convexity in $\boldsymbol{y}$ is straightforward, as the sum of a convex and a linear function (Boyd & Vandenberghe, 2004, Sections 3.2.1, 3.2.2, 3.3.1).

Property 3. This follows from the scaling property of the convex conjugate (Boyd & Vandenberghe, 2004, Section 3.3.2)

$$(t\Omega)^*(\boldsymbol{\theta}) = t\Omega^*(t^{-1}\boldsymbol{\theta})$$

Denoting $\boldsymbol{\eta}' = t^{-1}\boldsymbol{\eta}$, we have that

$$\begin{aligned} \ell_{t\Omega,\boldsymbol{A}}(\boldsymbol{\eta}, \boldsymbol{y}) &= (t\Omega_\triangle)^*(\boldsymbol{A}^\top \boldsymbol{\eta}) + t\Omega_\triangle(\boldsymbol{y}) - \boldsymbol{\eta}^\top \boldsymbol{A} \boldsymbol{y} \\ &= t\Omega^*_\triangle(\boldsymbol{A}^\top \boldsymbol{\eta}') + t\Omega_\triangle(\boldsymbol{y}) - \boldsymbol{\eta}^\top \boldsymbol{A} \boldsymbol{y} \\ &= t\big(\Omega^*_\triangle(\boldsymbol{A}^\top \boldsymbol{\eta}') + \Omega_\triangle(\boldsymbol{y}) - \boldsymbol{\eta}'^\top \boldsymbol{A} \boldsymbol{y}\big) = t\ell_{\Omega,\boldsymbol{A}}(t^{-1}\boldsymbol{\eta}, \boldsymbol{y}). \end{aligned}$$