Differentiation through solutions to optimization problems

Artyom Gadetsky

December 4th 2020

p(B|A)yesgroup.ru

HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY

Convex Constrained Optimization Problems

$$\tilde{x}(\theta) = \begin{cases} \arg\min_{x} f_0(x, \theta) \\ f(x, \theta) \leq 0 \\ h(x, \theta) = 0 \end{cases}$$

Convex Constrained Optimization Problems

$$\tilde{x}(\theta) = \begin{cases} \arg\min\limits_{x} f_0(x, \theta) \\ f(x, \theta) \le 0 \\ h(x, \theta) = 0 \end{cases}$$

$$\frac{\partial \tilde{x}(\theta)}{\partial \theta} \text{ - ?}$$

Naive method

$$\frac{\partial L(\tilde{x}(\theta))}{\partial \theta} = \frac{\partial \tilde{x}(\theta)}{\partial \theta}^{T} \frac{\partial L(\tilde{x}(\theta))}{\partial \tilde{x}(\theta)}$$

Naive method

$$\frac{\partial L(\tilde{x}(\theta))}{\partial \theta} = \frac{\partial \tilde{x}(\theta)}{\partial \theta}^T \frac{\partial L(\tilde{x}(\theta))}{\partial \tilde{x}(\theta)}$$

$$D_v f(x) = \frac{\partial f(x)}{\partial x} v = \lim_{\varepsilon \to 0} \frac{f(x + \varepsilon v) - f(x)}{\varepsilon}$$

Naive method

$$\frac{\partial L(\tilde{x}(\theta))}{\partial \theta} \approx \frac{\tilde{x}(\theta + \varepsilon \frac{\partial L(\tilde{x}(\theta))}{\partial \tilde{x}(\theta)}) - \tilde{x}(\theta)}{\varepsilon}$$

Naive method

$$\frac{\partial L(\tilde{x}(\theta))}{\partial \theta} \approx \frac{\tilde{x}(\theta + \varepsilon \frac{\partial L(\tilde{x}(\theta))}{\partial \tilde{x}(\theta)}) - \tilde{x}(\theta)}{\varepsilon}$$

$$\frac{\partial L(\tilde{x}(\theta))}{\partial \theta} \approx \frac{\tilde{x}(\theta + \varepsilon \frac{\partial L(\tilde{x}(\theta))}{\partial \tilde{x}(\theta)}) - \tilde{x}(\theta - \varepsilon \frac{\partial L(\tilde{x}(\theta))}{\partial \tilde{x}(\theta)})}{2\varepsilon}$$

Karush-Kuhn-Tucker Conditions

$$\tilde{x}(\theta) = \begin{cases} \arg\min_x f_0(x, \theta) \\ f(x, \theta) \leq 0 \\ h(x, \theta) = 0 \end{cases}$$

$$x \in \mathbb{R}^n, \;\; \theta \in \mathbb{R}^d$$

$$f_0 : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$$

$$f : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}^m$$

$$h : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}^p$$

Karush-Kuhn-Tucker Conditions

$$\tilde{x}(\theta) = \begin{cases} \arg\min\limits_{x} f_0(x, \theta) \\ f(x, \theta) \leq 0 \\ h(x, \theta) = 0 \end{cases}$$

$$x \in \mathbb{R}^n, \;\; \theta \in \mathbb{R}^d$$

$$f_0 : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$$

$$f : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}^m$$

$$h : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}^p$$

$$p^\star(\theta) = \inf\{f_0(x, \theta) | f(x, \theta) \preceq 0, h(x, \theta) = 0\}$$

Karush-Kuhn-Tucker Conditions

$$\tilde{x}(\theta) = \begin{cases} \arg\min_{x} f_0(x, \theta) \\ f(x, \theta) \leq 0 \\ h(x, \theta) = 0 \end{cases}$$

$$x \in \mathbb{R}^n, \;\; \theta \in \mathbb{R}^d$$

$$f_0 : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}$$

$$f : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}^m$$

$$h : \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}^p$$

$$p^\star(\theta) = \inf\{f_0(x, \theta) | f(x, \theta) \preceq 0, h(x, \theta) = 0\}$$

$$S(\theta) = \{x \mid f(x, \theta) \preceq 0, h(x, \theta) = 0, f_0(x, \theta) = p^\star(\theta)\}$$

Karush-Kuhn-Tucker Conditions

$$L(x, \lambda, \nu, \theta) = f_0(x, \theta) + \lambda^T f(x, \theta) + \nu^T h(x, \theta)$$

Karush-Kuhn-Tucker Conditions

$$L(x, \lambda, \nu, \theta) = f_0(x, \theta) + \lambda^T f(x, \theta) + \nu^T h(x, \theta)$$

$$\tilde{x}(\theta) \in S(\theta) \iff \exists(\tilde{\lambda}, \tilde{\nu}) : \quad \begin{aligned} f(\tilde{x}, \theta) &\preceq 0 \\ h(\tilde{x}, \theta) &= 0, \\ \tilde{\lambda}_i &\geq 0, \quad i = 1, \ldots, m \\ \tilde{\lambda}_i f_i(\tilde{x}, \theta) &= 0, \quad i = 1, \ldots, m \\ \nabla_x L(\tilde{x}, \tilde{\lambda}, \tilde{\nu}, \theta) &= 0. \end{aligned}$$

Karush-Kuhn-Tucker Conditions

$$\text{If } G = \{i | \tilde{\lambda}_i = 0 \text{ and } f_i(\tilde{x}, \theta) = 0\} = \emptyset \text{ then}$$

$$\tilde{x}(\theta) \in S(\theta) \iff \exists (\tilde{\lambda}, \tilde{\nu}) : \quad \begin{aligned} h(\tilde{x}, \theta) &= 0, \\ \tilde{\lambda}_i &\geq 0, \quad i = 1, \ldots, m \\ \tilde{\lambda}_i f_i(\tilde{x}, \theta) &= 0, \quad i = 1, \ldots, m \\ \nabla_x L(\tilde{x}, \tilde{\lambda}, \tilde{\nu}, \theta) &= 0. \end{aligned}$$

Idea

Denote $\tilde{r} = (\tilde{x}, \tilde{\lambda}, \tilde{\nu})$ then we have:

$$
\begin{aligned}
h(\tilde{x}, \theta) &= 0, \\
\tilde{\lambda}_i f_i(\tilde{x}, \theta) &= 0, \quad i = 1, \ldots, m \\
\nabla_x L(\tilde{x}, \tilde{\lambda}, \tilde{\nu}, \theta) &= 0.
\end{aligned}
\iff g(\tilde{r}, \theta) = 0
$$

Idea

Denote $\tilde{r} = (\tilde{x}, \tilde{\lambda}, \tilde{\nu})$ then we have:

$$\begin{aligned} h(\tilde{x}, \theta) &= 0, \\ \tilde{\lambda}_i f_i(\tilde{x}, \theta) &= 0, \quad i = 1, \ldots, m \\ \nabla_x L(\tilde{x}, \tilde{\lambda}, \tilde{\nu}, \theta) &= 0. \end{aligned} \iff g(\tilde{r}, \theta) = 0$$

Idea

Denote $\tilde{r} = (\tilde{x}, \tilde{\lambda}, \tilde{\nu})$ then we have:

$$\begin{aligned} h(\tilde{x}, \theta) &= 0, \\ \tilde{\lambda}_i f_i(\tilde{x}, \theta) &= 0, \quad i = 1, \ldots, m \\ \nabla_x L(\tilde{x}, \tilde{\lambda}, \tilde{\nu}, \theta) &= 0. \end{aligned} \iff g(\tilde{r}, \theta) = 0$$

Applying Implicit Function Theorem to KKT

$$\frac{\partial \tilde{r}(\theta)}{\partial \theta} = -\left(\frac{\partial g}{\partial \tilde{r}}\right)^{-1} \frac{\partial g}{\partial \theta}$$

Applying Implicit Function Theorem to KKT

$$\frac{\partial \tilde{r}(\theta)}{\partial \theta} = -(\frac{\partial g}{\partial \tilde{r}})^{-1}\frac{\partial g}{\partial \theta}$$

$$\frac{\partial L}{\partial \theta} = \frac{\partial \tilde{r}(\theta)}{\partial \theta}^T \frac{\partial L}{\partial \tilde{r}} = -(\frac{\partial g}{\partial \theta})^T(\frac{\partial g}{\partial \tilde{r}})^{-T}\frac{\partial L}{\partial \tilde{r}}$$

Applying Implicit Function Theorem to KKT

Denote $d_r = -(\frac{\partial g}{\partial \tilde{r}})^{-T} \frac{\partial L}{\partial \tilde{r}}$ then we have:

$$d_\theta L = \frac{\partial L}{\partial \theta}^T \partial \theta = d_r^T \frac{\partial g}{\partial \theta} \partial \theta = \sum_i \frac{\partial L}{\partial \theta_i}^T \partial \theta_i$$

OptNet: Quadratic Programs

$$\begin{aligned} \underset{z}{\text{minimize}} \quad & \frac{1}{2} z^T Q z + q^T z \\ \text{subject to} \quad & Az = b,\ Gz \leq h \end{aligned}$$

$$z \in \mathbb{R}^n$$

$$\theta = \{Q, q, A, b, G, h\} = \{S_+^n, \mathbb{R}^n, \mathbb{R}^{m \times n}, \mathbb{R}^m, \mathbb{R}^{p \times n}, \mathbb{R}^p\}$$

OptNet: Quadratic Programs

$$L(z, \nu, \lambda) = \frac{1}{2} z^T Q z + q^T z + \nu^T (Az - b) + \lambda^T (Gz - h)$$

OptNet: Quadratic Programs

$$L(z, \nu, \lambda) = \frac{1}{2} z^T Q z + q^T z + \nu^T (Az - b) + \lambda^T (Gz - h)$$

$$\begin{aligned} Qz^\star + q + A^T \nu^\star + G^T \lambda^\star &= 0 \\ Az^\star - b &= 0 \\ D(\lambda^\star)(Gz^\star - h) &= 0 \end{aligned} \iff g(\tilde{r}, \theta) = 0$$

OptNet: Quadratic Programs

$$L(z, \nu, \lambda) = \frac{1}{2} z^T Q z + q^T z + \nu^T (Az - b) + \lambda^T (Gz - h)$$

$$\begin{aligned} Qz^\star + q + A^T \nu^\star + G^T \lambda^\star &= 0 \\ Az^\star - b &= 0 \\ D(\lambda^\star)(Gz^\star - h) &= 0 \end{aligned} \iff g(\tilde{r}, \theta) = 0$$

OptNet: Quadratic Programs

$$\partial g(r,\theta) = \frac{\partial g}{\partial r}\partial r + \frac{\partial g}{\partial \theta}\partial \theta = 0 \iff \frac{\partial g}{\partial r}\partial r = -\frac{\partial g}{\partial \theta}\partial \theta$$

$$\Updownarrow$$

$$\begin{bmatrix} Q & G^T & A^T \\ D(\lambda^\star)G & D(Gz^\star - h) & 0 \\ A & 0 & 0 \end{bmatrix} \partial r = - \begin{bmatrix} \mathsf{d}Qz^\star + \mathsf{d}q + \mathsf{d}G^T\lambda^\star + \mathsf{d}A^T\nu^\star \\ D(\lambda^\star)\mathsf{d}Gz^\star - D(\lambda^\star)\mathsf{d}h \\ \mathsf{d}Az^\star - \mathsf{d}b \end{bmatrix}$$

OptNet: Quadratic Programs

$$d_r = -\left(\frac{\partial g}{\partial r}\right)^{-T} \frac{\partial l}{\partial r} = -\begin{bmatrix} Q & G^T D(\lambda^\star) & A^T \\ G & D(Gz^\star - h) & 0 \\ A & 0 & 0 \end{bmatrix}^{-1} \begin{bmatrix} \left(\frac{\partial \ell}{\partial z^\star}\right)^T \\ 0 \\ 0 \end{bmatrix}$$

OptNet: Quadratic Programs

$$d_\theta l = d_r^T \frac{\partial g}{\partial \theta} \partial \theta \quad = \mathrm{Tr}((\nabla_Q l)^T \partial Q) + \mathrm{Tr}((\nabla_A l)^T \partial A) + \mathrm{Tr}((\nabla_G l)^T \partial G) + $$
$$+ (\nabla_q l)^T \partial q + (\nabla_b l)^T \partial b + (\nabla_h l)^T \partial h$$

$$\Updownarrow$$

$$\begin{aligned}
\nabla_Q \ell &= \frac{1}{2}(d_z z^T + z d_z^T) & \nabla_q \ell &= d_z \\
\nabla_A \ell &= d_\nu z^T + \nu d_z^T & \nabla_b \ell &= -d_\nu \\
\nabla_G \ell &= D(\lambda^\star)(d_\lambda z^T + \lambda d_z^T) & \nabla_h \ell &= -D(\lambda^\star) d_\lambda
\end{aligned}$$

Disciplined Convex Programming (DCP)

$$\tilde{x} = \begin{cases} \arg\min_{x} f_0(x) \\ f(x) \leq 0 \\ h(x) = 0 \end{cases}$$

Disciplined Convex Programming (DCP)

$$\text{DCP} = (\mathcal{A}, S, E, \mathcal{R})$$

$$\begin{array}{lll} \mathcal{A} & - & \text{atom functions} \\ S & - & \text{"start" symbol} \\ E & - & \text{"end" symbol} \\ \mathcal{R} & - & \text{composition rule} \end{array}$$

$$\tilde{x} = \begin{cases} \arg\min_x f_0(x) \\ f(x) \leq 0 \\ h(x) = 0 \end{cases}$$

$$\mathcal{A} = (\{\text{convex}\} \cup \{\text{concave}\} \cup \{\text{affine}\}) \cap \{\text{monotone}\}$$

Composition theorem for DCP

$$I_1 \subseteq \{1, \ldots, k\}$$
$$I_2 \subseteq \{1, \ldots, k\}$$

Composition theorem for DCP

$$h(y) : \mathbb{R}^k \to \mathbb{R} \text{ - convex } + \begin{cases} h(y_{I_1}) \text{ - non-decreasing} \\ h(y_{I_2}) \text{ - non-increasing} \end{cases}$$

$$I_1 \subseteq \{1, \ldots, k\}$$
$$I_2 \subseteq \{1, \ldots, k\}$$

Composition theorem for DCP

$$h(y) : \mathbb{R}^k \to \mathbb{R} \text{ - convex } + \begin{cases} h(y_{I_1}) \text{ - non-decreasing} \\ h(y_{I_2}) \text{ - non-increasing} \end{cases}$$

$$g_i(x) : \mathbb{R}^n \to \mathbb{R} \text{ - } \begin{cases} \text{convex } \forall i \in I_1 \\ \text{concave } \forall i \in I_2 \\ \text{affine } \forall i \in (I_1 \cap I_2)^c \end{cases}$$

$$I_1 \subseteq \{1, \ldots, k\}$$

$$I_2 \subseteq \{1, \ldots, k\}$$

Composition theorem for DCP

$$h(y) : \mathbb{R}^k \to \mathbb{R} \text{ - convex } + \begin{cases} h(y_{I_1}) \text{ - non-decreasing} \\ h(y_{I_2}) \text{ - non-increasing} \end{cases}$$

$$g_i(x) : \mathbb{R}^n \to \mathbb{R} \text{ - } \begin{cases} \text{convex } \forall i \in I_1 \\ \text{concave } \forall i \in I_2 \\ \text{affine } \forall i \in (I_1 \cap I_2)^c \end{cases}$$

$$I_1 \subseteq \{1, \ldots, k\}$$
$$I_2 \subseteq \{1, \ldots, k\}$$

$$\Longrightarrow \; f(x) = h(g(x)) \text{ - convex}$$

CVXPY pipeline

$$x^*(\theta) = \begin{cases} \arg\min_{x} f_0(x, \theta) \\ f(x, \theta) \le 0 \\ h(x, \theta) = 0 \end{cases}$$

CVXPY pipeline

$$x^*(\theta) = \begin{cases} \arg\min_x f_0(x, \theta) \\ f(x, \theta) \leq 0 \\ h(x, \theta) = 0 \end{cases} \xrightarrow[\text{Form}]{\text{Canon}} \begin{cases} \arg\min_{\tilde{x}} c^T \tilde{x} \\ b - A\tilde{x} \in \mathcal{K} \end{cases}$$

CVXPY pipeline

$$x^*(\theta) = \begin{cases} \arg\min_x f_0(x,\theta) \\ f(x,\theta) \leq 0 \\ h(x,\theta) = 0 \end{cases} \xrightarrow[\text{Form}]{\text{Canon}} \begin{cases} \arg\min_{\tilde{x}} c^T \tilde{x} \\ b - A\tilde{x} \in \mathcal{K} \end{cases} \quad \xrightarrow[\text{Solver}]{\text{Conic}}$$

$$\xrightarrow[\text{Solver}]{\text{Conic}} \tilde{x}(A, b, c)$$

CVXPY pipeline

$$x^*(\theta) = \begin{cases} \arg\min_x f_0(x,\theta) \\ f(x,\theta) \le 0 \\ h(x,\theta) = 0 \end{cases} \xrightarrow[\text{Form}]{\text{Canon}} \begin{cases} \arg\min_{\tilde{x}} c^T \tilde{x} \\ b - A\tilde{x} \in \mathcal{K} \end{cases} \xrightarrow[\text{Solver}]{\text{Conic}}$$

$$\xrightarrow[\text{Solver}]{\text{Conic}} \tilde{x}(A,b,c) \xrightarrow[\text{Retrieval}]{\text{Solution}} x^*(\theta)$$

CVXPY pipeline

$$x^*(\theta) = \begin{cases} \arg\min_x f_0(x, \theta) \\ f(x, \theta) \leq 0 \\ h(x, \theta) = 0 \end{cases} \xrightarrow[\text{Form}]{\text{Canon}} \begin{cases} \arg\min_{\tilde{x}} c^T \tilde{x} \\ b - A\tilde{x} \in \mathcal{K} \end{cases} \xrightarrow[\text{Solver}]{\text{Conic}}$$

$$\xrightarrow[\text{Solver}]{\text{Conic}} \tilde{x}(A, b, c) \xrightarrow[\text{Retrieval}]{\text{Solution}} x^*(\theta)$$

$$\frac{\partial L(x^*(\theta))}{\partial \theta} = \frac{\partial x^*(\theta)}{\partial \theta}^T \frac{\partial L}{\partial x^* \theta}$$

CVXPY pipeline

$$x^*(\theta) = \begin{cases} \arg\min\limits_x f_0(x,\theta) \\ f(x,\theta) \leq 0 \\ h(x,\theta) = 0 \end{cases} \xrightarrow[\text{Form}]{\text{Canon}} \begin{cases} \arg\min\limits_{\tilde{x}} c^T \tilde{x} \\ b - A\tilde{x} \in \mathcal{K} \end{cases} \xrightarrow[\text{Solver}]{\text{Conic}}$$

$$\xrightarrow[\text{Solver}]{\text{Conic}} \tilde{x}(A,b,c) \xrightarrow[\text{Retrieval}]{\text{Solution}} x^*(\theta)$$

$$\frac{\partial L(x^*(\theta))}{\partial \theta} = \frac{\partial x^*(\theta)}{\partial \theta}^T \frac{\partial L}{\partial x^* \theta} \qquad \frac{\partial x^*(\theta)}{\partial \theta}^T = \frac{\partial C}{\partial \theta}^T \frac{\partial \tilde{x}}{\partial (A,b,c)}^T \frac{\partial R}{\partial \tilde{x}}^T$$

Convex Differentiable Optimization Layers

$$x^*(\theta) = \begin{cases} \arg\min_x f_0(x,\theta) \\ f(x,\theta) \le 0 \\ h(x,\theta) = 0 \end{cases} \xrightarrow[\text{Form}]{\text{Canon}} \begin{cases} \arg\min_{\tilde{x}} c^T \tilde{x} \\ b - A\tilde{x} \in \mathcal{K} \end{cases} \xrightarrow[\text{Solver}]{\text{Conic}}$$

$$\xrightarrow[\text{Solver}]{\text{Conic}} \tilde{x}(A, b, c) \xrightarrow[\text{Retrieval}]{\text{Solution}} x^*(\theta)$$

$$x^*(\theta) = R(\tilde{x}(C(\theta))) \qquad R, C \text{ - affine}$$

Disciplined Parametrized Programming (DPP)

DPP $\subset$ DCP such that:

$\cdot$parameters are classified as affine

$\cdot\phi_{\text{prod}}(x, y) = xy$ is affine if:

- $\cdot x$ or y is constant (parameter-free and variable-free)
- $\cdot x$ is parameter-affine and y is parameter-free or vice-versa

Example of DPP

$$\begin{cases} \min_x \|Fx - g\|_2 + \lambda\|x\|_2 \\ x \geq 0 \end{cases}$$

$$x \in \mathbb{R}^n, F \in \mathbb{R}^{n\times m}, g \in \mathbb{R}^m$$

$\phi_{\text{prod}}(F, x) \;=\; Fx$ is affine

$Fx - g$ is affine

$\|Fx - g\|_2$ is convex

$\phi_{\text{prod}}(\lambda, \|x\|_2)$ is convex

$$\mathcal{A} = \{\|\cdot\|_2, \text{product}, \text{negation}, \text{sum}\}$$

Example of Canonicalization

$$\begin{cases} \min_x \|Fx - g\|_2 + \lambda\|x\|_2 \\ x \geq 0 \end{cases}$$

$$x \in \mathbb{R}^n, F \in \mathbb{R}^{n\times m}, g \in \mathbb{R}^m$$

Example of Canonicalization

$$\begin{cases} \min\limits_{x} \|Fx - g\|_2 + \lambda\|x\|_2 \\ x \geq 0 \end{cases} \quad \xrightarrow{\text{C}} \quad \begin{cases} \min\limits_{t_1, t_2, x} t_1 + \lambda t_2 \\ (t_1, Fx - g) \in \mathcal{Q}_{m+1} \\ (t_2, x) \in \mathcal{Q}_{n+1} \\ x \in \mathbb{R}^n_+ \end{cases}$$

$$x \in \mathbb{R}^n, F \in \mathbb{R}^{n \times m}, g \in \mathbb{R}^m$$

$$A = \left[\begin{array}{ccc} -1 & & \\ & & -F \\ \hline & -1 & \\ & & -I \\ \hline & & -I \end{array}\right], \quad b = \begin{bmatrix} 0 \\ -g \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad c = \begin{bmatrix} 1 \\ \lambda \\ 0 \end{bmatrix}, \quad \mathcal{K} = \mathcal{Q}_{m+1} \times \mathcal{Q}_{n+1} \times \mathbf{R}^n_+$$

Example of Canonicalization

$$
\begin{cases}
\min\limits_{t_1, t_2, x} \; t_1 + \lambda t_2 \\
(t_1, Fx - g) \in \mathcal{Q}_{m+1} \\
(t_2, x) \in \mathcal{Q}_{n+1} \\
x \in \mathbb{R}^n_+
\end{cases}
\iff
\begin{cases}
\min\limits_{\tilde{x}} c^T \tilde{x} \\
b - A\tilde{x} \in \mathcal{K}
\end{cases}
$$

$$\tilde{x} = (t_1, t_2, x)$$

$$
A = \left[\begin{array}{ccc}
-1 & & \\
 & & -F \\
\hline
 & -1 & \\
 & & -I \\
\hline
 & & -I
\end{array}\right], \quad
b = \begin{bmatrix} 0 \\ -g \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad
c = \begin{bmatrix} 1 \\ \lambda \\ 0 \end{bmatrix}, \quad
\mathcal{K} = \mathcal{Q}_{m+1} \times \mathcal{Q}_{n+1} \times \mathbf{R}^n_+
$$

Applying IFT for canonical form

$$\tilde{x}(A, b, c) = \begin{cases} \arg\min_{\tilde{x}} c^T \tilde{x} \\ b - A\tilde{x} \in \mathcal{K} \end{cases}$$

$$\frac{\partial \tilde{x}(A, b, c)}{\partial (A, b, c)} = \mathrm{IFT}(A, b, c, \tilde{x})$$

Coming back to the original solution

After canonicalization we get:

$$\tilde{x} = (x, s)$$

Thus: $R(\tilde{x}^*) = x^*$ (slicing)

Putting it all together

$$x^*(\theta) = \begin{cases} \arg\min_x f_0(x,\theta) \\ f(x,\theta) \leq 0 \\ h(x,\theta) = 0 \end{cases} \xrightarrow[\text{Form}]{\text{Canon}} \begin{cases} \arg\min_{\tilde{x}} c^T \tilde{x} \\ b - A\tilde{x} \in \mathcal{K} \end{cases} \xrightarrow[\text{Solver}]{\text{Conic}}$$

$$\xrightarrow[\text{Solver}]{\text{Conic}} \tilde{x}(A,b,c) \xrightarrow[\text{Retrieval}]{\text{Solution}} x^*(\theta)$$

$$\frac{\partial L(x^*(\theta))}{\partial \theta} = \frac{\partial x^*(\theta)}{\partial \theta}^T \frac{\partial L}{\partial x^* \theta} \qquad \frac{\partial x^*(\theta)}{\partial \theta}^T = \frac{\partial C}{\partial \theta}^T \frac{\partial \tilde{x}}{\partial (A,b,c)}^T \frac{\partial R}{\partial \tilde{x}}^T$$

Example: Stochastic Softmax Tricks

Definition 1. *Given a non-empty, convex independent, finite set* $\mathcal{X} \subseteq \mathbb{R}^n$ *and a random utility* U *whose distribution is parameterized by* $\theta \in \mathbb{R}^m$, *a* stochastic argmax trick *for* X *is the linear program,*

$$X = \arg\max_{x \in \mathcal{X}} U^T x.$$

Example: Stochastic Softmax Tricks

Definition 1. *Given a non-empty, convex independent, finite set $\mathcal{X} \subseteq \mathbb{R}^n$ and a random utility U whose distribution is parameterized by $\theta \in \mathbb{R}^m$, a* stochastic argmax trick *for X is the linear program,*

$$X = \arg\max_{x \in \mathcal{X}} U^T x.$$

Definition 2. *Given a stochastic argmax trick $(\mathcal{X}, U)$ where $P := \mathrm{conv}(\mathcal{X})$ and a proper, closed, strongly convex function $f : \mathbb{R}^n \to \{\mathbb{R}, \infty\}$ whose domain contains the relative interior of P, a* stochastic softmax trick *for X at temperature $t > 0$ is the convex program,*

$$X_t = \arg\max_{x \in P} U^T x - t f(x)$$

Naive approximation

$$\frac{\partial L(X_t)}{\partial U} \approx \frac{X_t(U + \varepsilon \frac{\partial L(X_t)}{\partial X_t}) - X_t(U)}{\varepsilon}$$

$$\frac{\partial L(X_t)}{\partial U} \approx \frac{X_t(U + \varepsilon \frac{\partial L(X_t)}{\partial X_t}) - X_t(U - \varepsilon \frac{\partial L(X_t)}{\partial X_t})}{2\varepsilon}$$

References

- S. Barratt. *On the differentiability of the solution to convex optimization problems*. 2018. arXiv: 1804.05098.
- Brandon Amos and Zico Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *Proc. Intl. Conf. on Machine Learning (ICML)*, 2017.
- Agrawal, A., Amos, B., Barratt, S., Boyd, S., Diamond, S., and Kolter, J. Z. Differentiable convex optimization layers. In *Advances in neural information processing systems*, pp. 9562– 9574, 2019a.
- What is the Gradient of a Scalar Function of a Symmetric Matrix? Shriram Srinivasan and Nishant Panda. 2019: arXiv: 1911.06491
- Mattingley, Jacob, and Stephen Boyd. "CVXGEN: A code generator for embedded convex optimization." Optimization and Engineering13.1 (2012): 1-27.
- Max B. Paulus, Dami Choi, Daniel Tarlow, Andreas Krause, and Chris J. Maddison. Gradient estimation with stochastic softmax tricks. ArXiv, abs/2006.08063, 2020.